# Policy Change and Public Opinion: Measuring Shifting Political Sentiment with Social Media Data

Nicholas Adams-Cohen

January 2020

**Abstract**

This paper uses Twitter data and machine-learning methods to analyze the causal impact of the Supreme Court's legalization of same-sex marriage at the federal level in the United States on political sentiment and discourse towards gay rights. In relying on social media text data, this project constructs a large dataset of expressed political opinions in the short time frame before and after the *Obergefell v. Hodges* decision. Due to the variation in state laws regarding the legality of same-sex marriage prior to the Supreme Court's decision, I use a difference-in-difference estimator to show that, in those states where the Court's ruling produced a policy change, there was relatively more negative movement in public opinion towards same-sex marriage and gay rights issues as compared to other states. This confirms previous studies that show Supreme Court decisions polarize public opinion in the short-term, extends previous results by demonstrating opinion becomes relatively more negative in states where policy is overturned, and demonstrates how to use social media data to engage in causal analyses.

**Keywords:** Public Opinion, Political Communication, Text Analysis, Sentiment Analysis, Supreme Court

# Introduction

Researchers are divided over how the Supreme Court impacts American public opinion. One group of scholars argue that the public moves with the Justices' rulings, garnering consensus through strength of argument and the legitimacy of the courts (Lerner, 1967). Another camp argues that the public becomes further polarized after ruling on a divisive issue, with those inclined to agree with the Justices becoming more adamant in their support and those predisposed to disagreement becoming further entrenched in their opposition (Franklin & Kosaki, 1989). While these studies consider the ways in which public opinion changes temporally in the wake of a Supreme Court decision, few analyze the state-by-state reactions to federal rulings, critical if a decision aligns with one state's existing legal framework but overturns another's. This paper extends previous research into Supreme Court rulings and public opinion by incorporating a state-level analysis in studying the *Obergefell v. Hodges* decision and the federal legalization of same-sex marriage in the United States.

In June of 2015 in a 5-4 ruling, the Supreme Court held that the right to marry was a "fundamental right" under the Due Process Clause of the Fourteenth Amendment, instantaneously overturning same-sex marriage bans in thirteen states.[1] This decision represents the most recent in a long-line of monumental cases in which the Court made a ruling on a divisive social issue. Exploiting this variation in state laws regarding the legality of same-sex marriage, I use a difference-in-difference estimator to identify the *causal impact* of a *policy change* on the expression of sentiment towards same-sex marriage.[2] I find this impact to be negative, indicating a less positive response by those in the affected states, even when controlling for potentially relevant demographic variables and party identification.

While many studies use polling or survey data to measure the shift in public opinion before and after landmark decisions (e.g. Franklin & Kosaki, 1989; Johnson & Martin, 1998; Hanley, Salamone, & Wright, 2012; Christenson & Glick, 2015), this paper uniquely investigates these issues by using a subset of Twitter messages regarding same-sex marriage and gay rights issues.

By implementing machine learning methodologies to extract measures of sentiment from a large collection of tweets before and after the Supreme Court decision, I analyze a finely-grained dataset that allows for new insights into the short-term dynamics between Supreme Court decisions and public opinion. Studying this relationship is critical to discovering whether or not the Judiciary is able to guide public opinion in the direction of their opinion, or rather acts as a catalyst to further divide the public. The fine-grained temproal nature of social media data further allows me to examine the impact of the court's decision in the immediate weeks before and after the decision, examining the short-term reactions in these states to the court decision.

The paper proceeds as follows. First, I outline the relevant literature discussing the Supreme Court's impact on public opinion. Then, I describe the methodology I employ in this paper, including details on how I collect and analyze Twitter data. I then look at the aggregate change in opinion following that the Court's ruling before turning my analysis to a detailed investigation of the impact of policy change at the state-level. In this section, I conduct a difference-in-difference analysis, finding that the Supreme Court's decision engendered an increased negative reaction in those states where the ruling represented a change in state policy.

## The Supreme Court and Public Opinion

What constitutes the proper role of the Supreme Court in the United States is a long-standing question. In the opinion of Alexander Hamilton and the Federalists, the "independence of the judges may be an essential safeguard against the effects of occasional ill humors in the society" protecting against the "serious oppressions" of minority parties (Hamilton, Madison, & Jay, [2009] 1787-1788, p. 395-396). However, to counter the Anti-Federalists' arguments that an independent judiciary could wholly override the democratic process (Storing, 1981, p. 437-442), Hamilton further emphasized that the judiciary would be the "weakest" of the three branches of the Federal Government, without the "force nor will" to enforce its judgments (p. 392).

Without the ability to enforce its rulings, a number of scholars have argued that public opinion

constrains the Supreme Court (Hall, 2014). However, the Supreme Court's record demonstrates a number of instances where the court's rulings went against popular opinion, leading others to conclude the institution is counter-majoritarian in nature (Mishler & Sheehan, 1993). When decisions run counter to a majoritarian preference, scholars have argued the Court consciously recognized its role as "Republican Schoolmaster," using their judicial power to educate citizens and guide public opinion (Lerner, 1967).

Behind these arguments is the notion that, viewed as a popular and revered institution, the Supreme Court is able to directly influence public opinion in the direction of their decisions (Dahl, 1957; Casey, 1974; Mondak, 1994; Gibson & Caldeira, 2009).[3] The theory that the court lends legitimacy to their rulings in a way that moves public opinion in the direction of their decisions is termed the *Positive Response Hypothesis* (Franklin & Kosaki, 1989).

While there is a great deal of support for the *Positive Response Hypothesis* in experimental work (Clawson, Kegler, & Waltenburg, 2001; Mondak, 1994; Bartels & Mutz, 2009; Hoekstra, 2003), the theory does a poor job explaining empirical findings in a number of observational studies (Franklin & Kosaki, 1989; Nicholson & Hansford, 2014). *Roe v. Wade* represents a particularly important case study that refutes the *Positive Response Hypothesis*, as public opinion data show that before and after the ruling, aggregate support for abortion remained unchanged.

To address this empirical discrepancy, there have been a number of alternative theories describing how the public will respond to Supreme Court decisions. The *Structural Response Hypothesis* posits that, even if Supreme Court decisions fail to move aggregate public opinion in one direction, court decisions can still alter the "structure of opinion" – that is, the amount to which different groups "support and oppose a position and how intensely" (Franklin & Kosaki, 1989, p.753). Thus, a Supreme Court decision might cause ex-ante supporters of a position to become more favorable, while simultaneously causing ex-ante opponents to become more negative. In the aggregate, this would appear as no movement in overall public opinion, although in actuality the court was responsible in further polarizing public opinion.

Another alternative is the *backlash model*, predicting that Supreme Court rulings that change

policy will move aggregate public opinion away from the Justice's decision (Haider-Markel, 2007, 2010). In this model, Supreme Court decisions act as focusing events that lead to a "large, negative, and enduring shift in opinion against a policy or group" (Bishen, Hayes, Incantalupo, & Smith, 2016, p.626). We observe this backlash most acutely in the short-term, and it can eventually lead to long-term aggregate support to the Justice's decision (Ura, 2014). As I will detail below, I explicitly test these competing theories of how the might public respond to major Supreme Court decisions in a causal framework using social media data.

## Heterogeneous State Reactions

While there is extensive research and debate analyzing the impact of Supreme Court decisions on aggregate public opinion, much of the previous work does not explicitly test whether a shift in opinion is the same in the group of states where a ruling leads to a policy change, occurring whenever state and local policies contradict a Federal decision by the Supremacy Clause of the United States Constitution (U.S. Constitution Article VI, n.d., §2). My work addresses this gap in the literature by considering the consequences of Supreme Court decisions that nullify some state polices while leaving the policies of other states unchanged.[4]

The reason most earlier work does not consider the state-level reactions to Supreme Court rulings conditional on the state's existing legal framework is likely limitations in available data; with few comparable state-by-state surveys, researchers often often rely on national survey data (e.g. Marshall, 1989; Franklin & Kosaki, 1989; Johnson & Martin, 1998). However, given citizens in different parts of the country experience different policy consequences as a result of Supreme Court decisions, it seems natural to assume that different groups of states might have divergent reactions to the Justices' rulings.

To hypothesize how public opinion will move in states where the Supreme Court overturns policy, I consider the literature on public opinion towards Federalism. Survey data over the course of many years demonstrates that citizens consistently view their state governments more favorably than the Federal government (Kincaid & Cole, 2011, 2008, 2000). These "attitudes are sensi-

tive to respondents' affiliation with the party in power nationally" (p. 66), with members outside the standing President's party more likely to believe the federal government has too much power (2011). These opinions also vary by region, with citizens in southern states more likely to believe their state/province is not "treated with the respect it deserves in the federal system of government" (2008, p.g. 479). Given that the public tends to view state governments more favorably than the federal government, these studies suggest that when a Supreme Court decision goes against state-level policy, public opinion is prone to move *away* from the Justice's decision, a hypothesis I am able to test with my research design.

## Court Rulings and Opinion Towards Gay Rights

Prior to *Obergefell v. Hodges*, the Supreme Court ruled on a number of cases concerned with gay rights. While scholars analyzed the public response to these earlier cases, the empirical evidence across studies is mixed. Analyzing four separate gay rights cases, Stountengborough, Haider-Markel and Allen (2006) find public support moved in the direction of the court decision in one case, against the court decision in another, and remained unchanged for the remaining two cases.[5]

More recently, research analyzing the public's reaction to prominent Supreme Court cases expanding gay, including *Obergefell v. Hodges*, found little evidence that liberal decisions lead to a backlash against gay rights (Bishin, Hayes, Incantalupo, & Smith, 2016; A. R. Flores & Barclay, 2016; Kazyak & Stange, 2018).[6] Flores and Barclay (2016) further find that residents of states where the Court introduced same-sex marriage policy led to the greatest reduction in anti-gay attitudes. One potential reason these studies find little evidence of backlash is they utilize survey data, which often lags behind behind the date of a Supreme Court decision. This work may miss an initial, short-term backlash (Ura, 2014).

## Testing Hypotheses about the Public Response to *Obergefell v. Hodges*

Reviewing previous research allows me to come up with a number of predictions concerning the public's response to the Supreme Court's *Obergefell v. Hodges* ruling. Given the empirical support

for the *Structural Response Hypothesis*, I predict that, in the aggregate, the Supreme Court will polarize public opinion. In addition, given the literature on public attitudes towards Federalism, I believe in those states where the Supreme Court's decision resulted in a change in policy, there will be a more negative reaction towards the ruling as compared to other states in the short-term.

This allows me to develop two testable hypotheses:

H1. In the aggregate, the Supreme Court's ruling in the case of *Obergefell v. Hodges* will lead to further polarization towards attitudes on same-sex marriage and gay rights.

H2. In those states where the *Obergefell v. Hodges* ruling lead to a change in state-level policy, there will be a more negative reaction towards same-sex marriage and gay rights issues as compared to states where there was no change in policy.

# Twitter Data and Sentiment Scoring

Though I address the oft-discussed question of how Supreme Court decisions impact public opinion, I do so with a different methodology compared to past studies. Rather than relying on survey data, this paper utilizes machine learning sentiment analysis methodologies to obtain a measure of public opinion from Twitter messages. This section briefly describes how I obtain and proces this social media data and the strategies I used to quantify sentiment from raw text.

## Using Twitter Data to Study Opinion

While survey data is far and away the most popular source of data in studying public opinion, it is nearly impossible to collect for my present research question. First, in order to measure changes in public opinion before and after major Supreme Court cases, one needs to run comparable polls immediately before and after the Justice's reach their decision, a "limiting factor for all studies of Supreme Court influence on public opinion" (Brickman & Peterson, 2006, p.98). Second, studying the heterogenous impact of a Supreme Court decision requires strong state samples, with many

national surveys failing to report state-by-state results, given the margin-of-error for smaller states can be problematic for inference (Silver, 2016).[7] While researchers developed several techniques to estimate state samples from national survey data, including disaggregation (Erikson, Wright, & McIver, 1993) and mulitivelel regression and poststratification (MRP) (Lax & Phillips, 2009a, 2009b), the additional necessity in finding comparable national surveys immediately before and after a ruling makes it difficult to use these techniques to study the short-term reactions to Supreme Court decisions.

Collecting messages on a site like Twitter is a potential way to circumvent these issues. Users send tweets in real-time, allowing for much finer-grained estimates of public opinion in comparison to with monthly (or even weekly) polls. Twitter data is also 'always-on,' making it possible to continuously collect information without needing to specify where and when to conduct a particular survey (Salganik, 2018, p. 21), allowing a researcher to study a wide range of unexpected events that alter might public sentiment and discourse. While an imperfect substitute to well collected polling data, many researchers have demonstrated how Twitter data can provide a strong signal of the public opinion (e.g. O'Connor, Balasubramanyan, & Routledge, 2010; McKelvey, DiGrazia, & Rojas, 2014; Beauchamp, 2017).

Of course, one of the major weaknesses of Twitter data is the fact that the population of American Twitter users is not a representative sample of the adult population in the United States. Research into the demographic makeup of Twitter users shows that populous American counties tend to be over represented (Mislove, Lehman, Ahn, Onnela, & Rosenquist, 2011), users are more likely to be younger and richer (Barberà & Rivero, 2015), and overall there is a liberal and pro-Democratic bias compared with the country as a whole (Mitchell & Hitlin, 2013). While a non-representative population is not a unique feature with social media data (non-response rates in polls produces a similarly difficult to correct bias (e.g. Groves & Peytcheva, 2008; Massey & Tourangeau, 2013; Desilver & Keeter, 2015)), it does make it difficult to claim high external validity in any study relying on Twitter data. That said, studying the population on Twitter does allow for research designs that maintain strong internal validity, allowing us to consider the comparative

statistics beyond the overall level of estimated effects. Given the lack of alternative polling data differentiated by state over the short time-frame around the *Obergefell v. Hodges* decision, Twitter data represents the best alternative to conduct a causal analysis.[8]

## Gathering Twitter Data

In order to utilize Twitter data to study changes in opinion concerning same-sex marriage, it is first necessary to filter through the vast quantity of Twitter data and obtain only the subset of messages where users discuss topics relating to gay marriage and rights. I accomplish this by utilizing the Twitter Streaming API, a tool that pulls any tweet that fits certain criteria in real-time.[9] To obtain all relevant tweets, I tracked the following set of words: **gay marriage, gay marriages, same-sex marriage, same-sex marriages, same sex marriage, same sex marriages, same-sex union, same-sex unions, same sex union, same sex unions, marriage equality, equal marriage.**[10] I pulled tweets containing one of these keywords from the Twitter Streaming API and placed into a MySQL data base with a Python script. This monitor ran from May 27, 2015 to August 24, 2015, collecting 5,996,741 total tweets.

For each tweet collected, several other pieces of relevant meta-data were captured, including the time-stamp a message was sent and the user's number of followers. When available, I also collect user profile data, such as the user's full name and location.[11]

As the goal of this project is to analyze sentiment within the United States, I focus on the subset of tweets with location data that can be mapped to a specific US state. I rely on self-reported locations to map users into US states. Specifically, I employ a large series of regular expressions with state names and the most populous American cities to map self-reported location data into a standardized state-coding scheme. In total, I mapped 1,028,151 messages to a specific state.[12]

In addition to analyzing location data, I examine the subset of users that choose to share their full name to predict demographic characteristics. Specifically, I use the `gender` package (Mullen, 2015) to link first names to gender and the `wru` package (Khanna & Imai, 2015) to link surnames with race. While it is impossible to perfectly predict gender and race based on names, these

packages are commonly employed in the literature, utilizing census data to predict these variables. In total, 481,487 messages were from users with names that could be linked to race and gender.

Finally, I classify a subset of users as either Republicans or Democrats. These labels come from data collected by Pablo Barberà (2013). Very briefly, Barberàs work takes advantage of follower networks to predict the likelihood an individual is a Republican or Democrat, with the estimation strategy relying on the logic that a Republican is more likely to follow other Republicans and Democrats are more likely to follow other Democrats. I was able to merge 184,042 my own Twitter data with users in Barberà's data, creating a subset of accounts with estimated party labels.

## Sentiment Scoring

After collecting a large set of Twitter data, I preprocess the raw text data in a way that made it possible to utilize various supervised sentiment scoring algorithms to measure.[13] Supervised training methods require a training set, a collection of messages annotated with true labels. As the goal of this project is to classify tweets based on sentiment, this involves building a training set of tweets labeled as *positive* or *negative*.

I use the crowd sourcing platform Mechanical Turk to obtain a set of hand-annoted tweets to build a binary sentiment classifier. Each tweet was labeled by three human coders, with the final label being the majority category.[14] In total, I collect a set of 626 negative and 1,778 positive unique tweets.d[15] In order to create a training set most representative of the corpus, I gather labels for the most-retweeted messages in my data collection period, with the 626 unique negative messages representing 44,031 total tweets, and the 1,778 unique positive messages representing 161,525 total tweets.

With this training set, I test a number of supervised classifiers, settling on random forest as the algorithm that leads to the best results (more details on this procedure can be found in the online appendix). Leaving out 10% of the training data for a test set, my final model specification has 81.74% accuracy and a Cohen's kappa coefficient of 0.40. Of the 1,028,151 tweets I map to a US state, I classify 182,031 as negative and 846,120 as positive. On acquiring this well-performing

estimate of sentiment in a carefully selected subset of tweets, I use these data as a measure of sentiment towards gay-rights issues before and after the Supreme Court's federal legalization of same-sex marriage.

# Aggregate Shift in Public Opinion

I begin by testing the *Structural Response Hypothesis* by replicating the analysis outlined in Franklin and Kosaki (1989). This model takes the form:

$$Y_i = \alpha_1 + \alpha_2 After_i + (\beta_{11} + \beta_{21} After_i) X_1 + (\beta_{12} + \beta_{22} After_i) X_2 + \ldots + (\beta_{1k} + \beta_{2k} After_i) X_k + \epsilon$$

Where $i$ indexes messages, $Y$ is a "positive" or "negative" classifier, $After$ is an indicator variable specifying whether the message was from before or after the Supreme Court ruling, $X$ represents covariates, and $\epsilon$ represents unobservables. To measure $Y$, I use the random forest classifier described in the previous section to label each message in my dataset as positive or negative, replacing Random Forset scores with hand-labeled Mechanical Turk results when available (the hand-annotated labels are closer to the ground truth). I code positive message as a one and negative message as a zero. I estimate the above equation with a probit model.

(Table 1)

To test the *Structural Response Hypothesis*, I run two models: a constrained model in which $\beta_{2k}$ is set to zero for all $K$ covariates, and an unconstrained model where these values are allowed to vary. If I reject the constrained model in favor of the unconstrained model, it demonstrates that the Supreme Court decision alters the *structure* of opinion. I run two pairs of models: a pair that only includes demographic variables and a pair that includes demographic variables and party labels. Table 1 contains the results of these tests.[16]

In both pairs of models, I reject the constrained model in favor of the unconstrained model at high levels of significance, which confirms my first hypothesis (H1). Of note is the fact that this

level of significance is much higher when including party fixed effects, as demonstrated by the much larger Chi-Squared value across models three and four. This pattern demonstrates that the polarizing impact of *Obergefell v. Hodges* was especially pronounced across party lines.

Overall, these results provide further evidence for the *Structural Response Hypothesis*, demonstrating that the Supreme Court polarizes aggregate public opinion. Confirming the core result of Franklin and Kosaki (1989) represents a good initial validation of the accuracy of my sentiment classifier.

# Impact of Policy Change

To test how the Supreme Court's ruling in *Obergefell v. Hodges* may have affected the expression of sentiment towards gay marriage for citizens in regions where the Supreme Court overturned state-level policy, I use a difference-in-difference estimator to identify a treatment effect. The difference-in-difference estimator works by differencing across the treated and untreated observations, as well as across time. This effectively differences out both the time-variant and time-invariant unobservables, allowing for a causal interpretation of the difference-in-difference coefficient.

However, this estimation technique is only useful if what occurred in the untreated set is a reasonable counter-factual for what might have happened in the treated set. Thus, in this setting, I assume the treated states would have had a similar reaction as their untreated counterparts if the Supreme Court decision *did not* lead to a top-down shift in state policy. Importantly, the level of sentiment can still differ greatly between the two sets of states: only the general time-trend must be the same, an assumption explored below.

When these assumptions hold, there is no need for a difference-in-difference estimation to include other covariates. However, as the parallel trends assumption is very difficult to test, I include a number of covariates that could reasonably explain the heterogeneous response to the Supreme Court ruling across each set of states.

The difference-in-difference regression takes the form:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 After_i + \beta_3 (D_i * After_i) + X_i + \epsilon$$

Where $i$ indexes messages, $Y$ is a "positive" or "negative" classifier (defined in the way described in the previous section), $D$ is a treatment indicator that takes the value of 1 if the user sent the tweet if the Supreme Court decision lead to a change in state policy, and $After$ is an indicator variable that takes on the value 1 if the user sent the tweet after the Supreme Court's decision on June 26, 2015. $X$ represents a set of potential control variables and $\epsilon$ represents unobservables. I run this regression with a linear probability model, as the assumptions of the difference-in-difference estimator require linearity in order to interpret the results causally.

The coefficient of interest in the above equation is $\beta_3$, which corresponds with the average change in the expression of positive sentiment in the treatment group before and after the Supreme Court decision, minus the change in sentiment over the same period of time in the untreated group. This difference-in-difference represents the change in sentiment caused by the treatment, in this case the change in sentiment that results from the Supreme Court overturning state-level policy.

In total, I consider five models, and present the results in Table 2. The first model is the baseline difference-in-difference models, with no added controls. The second model removes all tweets sent on June 26, the day of the Supreme Court decision, in order to look beyond the impact of individuals *only* tweeted on June 26 and no other point in the dataset. Model three includes gender and race fixed effects, model four includes partisan labels, model five includes both.

(Table 2)

In Table 2, I find a negative and statistically significant **Treated**×**After** coefficient across the first four model specifications, providing strong empirical evidence that the Supreme Court's decision lead to a more *negative* reaction in those states where the decision lead to a policy change. Thus, I find evidence for my second hypothesis (H2): the Supreme Court's decision caused short-term backlash against gay marriage and gay rights in those affected states.

In model five, I find a null result. While this model includes the most covariates, I severely restrict the dataset by only considering users with names that could be linked to race and gender and could be matched to party labels. In total, this limits me to less than 10% of the original data, and losing statistical power is one reason I may fail to recover the effect.

Overall, the highly significant and negative **Treated**$\times$**After** coefficient across the first four models demonstrates the relative backlash in the treated states after the Court decision. Thus, these models provide support that, when the Supreme Court overturns state policy, there is less relative support for the decision in affected states. While this may seem to contradict earlier results that demonstrate a positive response to the Justice's decision, this is most likely do to the very short time frame of my current study. Previous studies (A. R. Flores & Barclay, 2016; Kazyak & Stange, 2018) analyzing the publics' reaction to *Obergefell v. Hodges* use survey data that lag behind and ahead of a Supreme Court case, perhaps failing to identify a short-term backlash.

Turning to each model in detail, models one and two represent baseline models, with and without tweets from the day of the Supreme Court decision. Looking at the **After** coefficient across models 1 and 2, I find that ignoring the strong positive response immediately following the Justice's decision on June 26, there appears to be an overall backlash in the short-term. The **Treated** coefficient is also negative in all models that do not include party labels. In model three, I find that including party labels leads to a statistically significant and positive **Treated** coefficient, demonstrating that when controlling for partisanship, there was overall more positive messages in these states. Even in this model specification, however, **Treated*After** remains negative and significant, demonstrating that there was a relatively less positive response to the court ruling in the treated states. The large, negative, and highly significant **Republican** coefficients in models four and five is not surprising, as conservative groups (consisting of mostly Republicans) consistently respond negatively to policies that advance a gay-rights agenda.

## Parallel Trends Assumption

While these results do not definitively prove causality, they demonstrate that the Supreme Court overturning state policy is correlated with less positive sentiment towards gay marriage and gay rights issues. If the untreated states are a good counter-factual to the treated states, this correlation can be interpreted causally.

This requires me to consider the untreated states as a good counter-factual to what might have occurred in the treated states. Unfortunately, this assumption is impossible to test. That said, if I can demonstrate that the treated and untreated states had a parallel trend in expressed sentiment prior to June 26, I can argue that the untreated set is a good control group for the treated set.[17] To explore this parallel trend assumption, I graph the differences in the daily mean sentiment score for treated and untreated states over time. As these daily sentiment scores are volatile, I chart the seven-day moving average (and LOESS curve in blue) to better visualize the data. This visualization is found in Figure 1.

(Figure 1)

In Figure 1, I find that overall the parallel trends assumption seems to hold, as both the treated and untreated states have the similar overall trend in expressed sentiment prior to June 26. For the most part, treated states have lower sentiment scores than their untreated counterparts, though there are periods of time where the scores overlap. After the court decision there was a general widening in the gap between sentiment scores across the two sets of states, a gap driving the difference-in-difference results. This gap is especially pronounced around July 1 to July 20. Exploring messages from these days might elucidate why there was an increase in negative sentiment during these time-periods, although this investigation is beyond the scope of the present paper. Near the end of the period of analysis, the treated and untreated states once again begin to converge, indicating a possible mean-reversion. This again demonstrates that my finding could point towards short-term backlash, with an eventual positive response later in the time trend.

# Conclusion

In this project, I bring a new perspective to the long-standing debate on how the Supreme Court impacts public opinion. In the landmark case *Obergefell v. Hodges*, the Supreme Court definitively ruled that same-sex marriage was a "fundamental right," conferring the right to marry for same-sex couples across the United States. As same-sex marriage is a divisive social issue, previous studies theorize this Supreme Court decision would cause opinion to be further polarized across the American public.

However, few earlier studies explicitly consider Supreme Court decisions' heterogeneous impact on different groups of states – with varying pre-existing legal conditions, a court ruling might overturn certain state policies while leaving other policies unchanged. Such was the case in *Obergefell v. Hodges*, with only thirteen of the fifty states having policy overturned in the wake of the Justice's decision. I study this event in a causal inference framework with a difference-in-difference estimator, finding that overturning state-level policy led to a relatively more negative reaction towards the decision by citizens in those affected states.

I engage in this analysis with a novel dataset: rather than conducting my study with public opinion polling data, I utilize machine learning methods to classify a large set of political tweets as *positive* or *negative* with a high degree of accuracy. These data allow me to track the expressed sentiment of gay rights issues in a short time frame, making it possible to detect shifts in sentiment immediately before and after the Supreme Court's decision. While social media data have their own set of potential issues, relying on Twitter allows me to construct a large dataset with coverage across the entire United States over the short period of time before and after the Court ruling, a necessary precondition in conducting a difference-in-difference analysis and allowing me to analyze the effect of the Supreme Court case in a causal framework.

This work represents a theoretical and methodological contribution to the literature on the Supreme Court's impact on public opinion. On the theory side, my work demonstrates that analyzing national-survey data without considering state samples is insufficient in understanding the

impact of Supreme Court decisions on public sentiment when those decisions have varied regional consequences. This work suggests that, when the Supreme Court overturns state policy, it leads to a relative short-term backlash against the Justice's decision. Future work should extend the data collection period to discover how this trend changes in the months and years after a decision is reached, as well as look at new court cases in different issue areas to establish this as a general finding.

On the methodological side, I demonstrate that combining sentiment analysis techniques with social media data grants a new perspective in analyzing public opinion. These data allow me to isolate the state-by-state reactions immediately following and preceding the *Obergefell v. Hodges* Supreme Court ruling, making it possible to analyze reactions to policy change in a causal inference framework, a unique contribution to this literature. In the future, gaining a better understanding of the demographic population on Twitter and improving the machine learning classification techniques will improve this methodology, allowing researchers to better understand and correct the bias in analyzing a non-representative sub-population.

# Endnotes

[1]This group of 13 states are: Arkansas, Georgia, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Ohio, South Dakota, Tennessee, and Texas.

[2]Sentiment, broadly defined, is an expression of an individual's "opinions, sentiments, evaluations, appraisals, attitudes, and emotions" towards a particular event, topic, or object (Liu, 2012). Public opinion refers to a citizen's feelings regarding an important political issue (Norrander & Wilcox, 2001). As the terms are closely linked, political sentiment and public opinion are used interchangeably in this work.

[3] While the popularity of the Supreme Court ebbs and flows over time (Caldeira, 1986), it is often shown to be perceived as more favorable than either the Legislative or Executive Branch (Cox, 1976; Marshall, 1989, p.g. 139-141).

[4]It is important to note that a number of the thirty-seven states that legalized same-sex marriage prior to the *Obergefell v. Hodges* decision did so only as the result of a state or district court ruling. While possible to assume citizens in these states would have the same reaction as the citizens in the thirteen states where *Obergefell v. Hodges* lead to a policy change, the *backlash model* theorizes citizens with direct exposure to focusing events are more likely to have a negative reaction to a policy (Hopkins, 2010). Therefore, even within this group of states, it is plausible that citizens in states where *Obergefell v. Hodges* lead to a policy change would have an increased negative reaction towards the ruling.

[5]The four cases studies were *Bowers v. Hardwick* (1986), *Romer v. Evans* (1996), *Boy Scouts of America v. Dale* (2000), and *Lawrence v. Texas* (2003).

[6]These cases include *United States v. Windsor (2013)*, which invalidated sections of the Defense of Marriage Act, and *Hollingsworth v. Perry (2013)*, which effectively legalized same-sex marriage in California.

[7]A good example of this limitation can be found when observing the Pew Research Center's (2016) report on changing attitudes towards gay marriage. While age, religion, party identification, race, and gender are among the reported covariates, state data are not provided.

[8]Flores (2017) uses this same research design to study whether anti-immigrant laws lead to a backlash against immigration related issues with Twitter data.

[9] A potential issue in utilizing data from Twitter's Streaming API is you do not get access to the full universe of messages. However, as there is no systemic pattern to which data are unavailable from the API, the bias this introduces is small when collecting a large dataset.

[10] While I specified these keywords to follow a single issue over time, in relying on a static list of keywords, I risk missing important phrases that developed dynamically during the data collection period (King, Lam, & Roberts,

2017). However, one advantage in using a static list of terms is I use the same criteria to select tweets during the entirety of my data collection period.

[11]A primary concern when collecting Twitter data is the potential incidence of 'bot' accounts – automated programs that perform a variety of actions on Twitter including sending messages, following other users, or retweeting messages (Jajodia, Wang, Gianvecchio, & Chu, 2012). While potentially problematic, an examination of the users in my data reveals little evidence that a large number of users are bots (see *Appendix F: Checking for Bot Accounts* for details). Based on this, I do not believe bots heavily bias my results.

[12]The most accurate form of location information in Twitter data are the GPS coordinates (geotags) users can elect to post with their tweets (Steinert-Threkeld, 2018, p.g. 7). However, as only only 2-3% of all tweets contain geotags (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013), this would not provide enough location tweets for me to engage in my analysis. In order to get a sense of how well my location coding scheme performs, I do analyze the subset of messages (1,990 in total) that are geotagged. For each of these messages, my mapping algorithm correctly classifies the user's state 91.1% of the time, providing a good robustness check for the mapping algorithm.

[13]Details of the preprocessing scheme can be found in *Appendix B: Preprocessing Text Data.*

[14]If all three Mechanical Turkers choose different categories, I dropped the data point. Only 5% of the tweets had no majority category, demonstrating high inter-coder reliability.

[15]While I choose to focus on a binary classifier given the stronger performance of these models on small training sets, I do consider a three-way classifier that incorporates neutral tags. These results can be found in Appendix D: *Neutral Tweets* and lead to even stronger substantive results.

[16]All regression tables are made with `stargazer` for R (Hlavac, 2015).

[17]To further test the parallel trends assumption, I restrict the untreated set to locations bordering the treated states in Appendix E: *Border State Analysis*. These results largely confirm my main findings.

# References


Barberà, P. (2013). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *American Behavioral Scientist*, *58*, 556-573.

Barberà, P., & Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, *33*, 712-729.

Bartels, B. L., & Mutz, D. C. (2009). Explaining processes of institutional opinion leadership. *The Journal of Politics*, *71*(1), 249-261.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, *61*(2), 490-503.

Benoit, K., & Nulty, P. (2016). quanteda: Quantitative analysis of textual data [Computer software manual]. Retrieved from `http://github.com/kbenoit/quanteda` (R package version 0.9.1-11)

Bishen, N. G., Hayes, T. J., Incantalupo, M. B., & Smith, C. A. (2016). Opinion backlash and public attitudes: Are political advances in gay rights counterproductive? *American Journal of Political Science*, *60*(3), 625–648.

Bishin, B. G., Hayes, T. J., Incantalupo, M. B., & Smith, A. (2016). Opinion backlash and public attitudes: Are political advances in gay rights counterproductive? *Political Research Quarterly*, *69*(1), 43–56.

Brickman, D., & Peterson, D. A. M. (2006). Public opinion reactionto events: Citizen response to multiple supreme court abortion decisions. *Political Behavior*, *28*(1), 87–112.

Caldeira, G. A. (1986). Neither the purse nor the sword: Dynamics of public confidence in the supreme court. *American Political Science Association*, *80*(4), 1209-1226.

Casey, G. (1974). The supreme court and myth: An empirical investigation. *Law & Soceity Review*, *8*(3), 385–420.

Christenson, D. P., & Glick, D. M. (2015). Issue-specific opinion change the supreme court and health care reform. *Public Opinion Quarterly*, *79*(4), 881-905.

Clawson, R. A., Kegler, E. R., & Waltenburg, E. N. (2001). The legitimacy-conferring authority of the U.S. supreme court: An experimental design. *American Politics Research*, *29*(6), 566-591.

Cox, A. (1976). *The role of the supreme court in american government*. Oxford University Press.

Dahl, R. (1957). Decision-making in democracy: The supreme court as national policy-maker. *Journal of Public Law*, *6*(2), 279–295.

Desilver, D., & Keeter, S. (2015). The challenges of polling when fewer people are available to be polled. *Pew Research Center*. Retrieved from `http://www.pewresearch.org`

Erikson, R. S., Wright, G. C., & McIver, J. P. (1993). *Statehouse democracy: Public opinion and policy in the american states*. Cambridge: Cambridge University Press.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Flores, A. R., & Barclay, S. (2016). Backlash, consensus, legitimacy, or polarization: The effect of same-sex marriage policy on mass attitudes. *Political Research Quarterly*, *69*(1), 43–56.

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, *123*(2), 333—384.

Franklin, C. H., & Kosaki, L. C. (1989). Republican schoolmaster: The u.s. supreme court, public opinion, and abortion. *American Political Science Review*, *83*(3), 751-71.

Gibson, J. L., & Caldeira, G. A. (2009). *Citizens, courts, and confirmations: Positivity theory and the judgments of the american people*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267-297.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public Opinion Quarterly*, *72*, 167-189.

Haider-Markel, D. P. (2007). Representation and backlash: The positive and negative influence of descriptive representation. *Legislative Studies Quarterly*, *32*(1), 107–133.

Haider-Markel, D. P. (2010). *Out and running: Gay and lesbian candidates, elections, and policy*

*representation*. Washington, DC: Georgetown University Press.

Hall, M. E. (2014). The semiconstrained court: Public opinion, the separation of powers, and the U.S. supreme court's fear of nonimplementation. *American Journal of Political Science*, *58*(2), 352-366.

Hamilton, A., Madison, J., & Jay, J. ([2009] 1787-1788). *The federalist papers* (I. Shapiro, Ed.). Yale University Press.

Hanley, J., Salamone, M., & Wright, M. (2012). Reviving the schoolmaster: Reevaluating public opinion in the wake of roe v. wade. *Political Research Quarterly*, *65*(2), 408-421.

Hlavac, M. (2015). stargazer: Well-formatted regression and summary statistics tables [Computer software manual]. Cambridge, USA. Retrieved from `http://CRAN.R-project.org/package=stargazer` (R package version 5.2)

Hoekstra, V. J. (2003). *Public reaction to supreme court decisions*. Cambridge University Press.

Hopkins, D. J. (2010). Politicized places: Explaining where and when immigrants provoke local opposition. *American Political Science Review*, *104*(1), 40–60.

Jajodia, S., Wang, H., Gianvecchio, S., & Chu, Z. (2012, 11). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*, 811-824.

Johnson, T. R., & Martin, A. D. (1998). The public's conditional response to supreme court decisions. , *92*(2), 299-309.

Kazyak, E., & Stange, M. (2018). Backlash or a positive response?: Public opinion of lgb issues after obergefell v. hodges. *Journal of Homosexuality*, *65*(14), 2028—2052.

Khanna, K., & Imai, K. (2015). Who are you? bayesian prediction of racial category using surname and geolocation [Computer software manual]. Retrieved from `https://github.com/ropensci/gender` (R package version 0.5.1)

Kincaid, J., & Cole, R. L. (2000). Public opinion and american federalism: Perspectives on taxes, spending, and trust: An acir update. *Publius*, *30*, 189-201.

Kincaid, J., & Cole, R. L. (2008). Public opinion on issues of federalism in 2007: A bush plus?

*Publius*, *38*, 469-487.

Kincaid, J., & Cole, R. L. (2011). Citizen attitudes towards issues of federalism in canada, mexico, and the united states. *Publius*, *41*, 53-75.

King, G., Lam, P., & Roberts, M. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*. Retrieved from `http://j.mp/2nxUa8N`

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26. Retrieved from `https://www.jstatsoft.org/v028/i05` doi: 10.18637/jss.v028.i05

Lax, J. R., & Phillips, J. H. (2009a). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, *103*(3), 367–186.

Lax, J. R., & Phillips, J. H. (2009b). How should we estimate public opinion in the states. *American Journal of Political Science*, *53*(1), 107–121.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, *18*(5).

Lerner, R. (1967). The supreme court as republican schoolmaster. *The Supreme Court Review*, *1967*, 127-180.

Liu, B. (2012). *Sentiment analysis and opinion mining* (G. Hirst, Ed.). Morgan & Claypool Publishers.

Marshall, T. R. (1989). *Public opinion and the supreme court*. Unwin Hyman.

Massey, D. S., & Tourangeau, R. (2013). Where do we go from here? nonresponse and social measurement. *ANNALS of the American Academy of Political and Social Science*, *645*(1), 222–236.

Maynard, D., & Greenwood, M. (2014, 01). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. *Proceedings of LREC*, 4238-4243.

McKelvey, K., DiGrazia, J., & Rojas, F. (2014). Twitter publics: how online political communities signaled electoral outcomes in the 2010 us house election. *Information, Communication*

*Society*, *17*(4), 490-503.

Mishler, W., & Sheehan, R. S. (1993). The supreme court as a countermajoritarian institution? the impact of public opinion on supreme court decisions. *American Political Science Review*, *87*(1), 87-101.

Mislove, A., Lehman, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Proceedings of the fifth international aaai conference on weblogs and social media* (pp. 554–557). AAAI Press.

Mitchell, A., & Hitlin, P. (2013). Twitter reaction to events often at odds with overall public opinion. *Pew Research Center*. Retrieved from `http://www.pewresearch.org`

Mondak, J. J. (1994). Policy legitimacy and the supreme court: The sources and contexts of legitimation. *Political Research Quarterly*, *47*(3), 675-692.

Mullen, L. (2015). gender: Predict gender from names using historical data [Computer software manual]. Retrieved from `https://github.com/ropensci/gender` (R package version 0.5.1)

Nicholson, S. P., & Hansford, T. G. (2014). Partisans in robes: Party cues and public acceptance of supreme court decisions. *American Journal of Political Science*, *58*(3), 620-636.

Norrander, B., & Wilcox, C. (2001). *Understanding public opinion, 2nd edition*. CQ Press.

O'Connor, B., Balasubramanyan, R., & Routledge, B. R. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the fourth international aaai conference on weblogs and social media.*

Pew Research Center. (2016, May). Changing attitudes on gay marriage. Retrieved from `http://www.pewforum.org`

Potts, C. (2011). Sentiment symposium tutorial [Computer software manual]. Retrieved from `http://sentiment.christopherpotts.net/index.html`

Salganik, M. J. (2018). *Bit by bit: Social research in the digital agae*. Princeton University Text.

Silver, N. (2016, September). Is a 50-state poll as good as 50 state polls? *FiveThirtyEight*. Retrieved from `http://fivethirtyeight.com`

Steinert-Threkeld, Z. C. (2018). *Twitter as data*. Cambridge University Press.

Storing, H. J. (1981). The complete anti-federalist. In (Vol. 2). The University of Chicago Press.

Stoutenborough, J. W., Haider-Markel, D. P., & Allen, M. D. (2006). Opinion backlash and public attitudes: Are political advances in gay rights counterproductive? *Political Research Quarterly*, *59*(3), 419–433.

Ura, J. D. (2014, January). Backlash and legitimation: Macro political responses to supreme court decisions. *American Journal of Political Science*, *58*, 110-126.

U.S. Constitution Article VI, S. . (n.d.).

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv:1703.03107*.

Wang, A. H. (2010). Detecting spam bots in online social networking sites: A machine learning approach. In *Ifip annual conference on data and applications security and privacy* (pp. 335–342).

# Tables

Table 1: Structural Response Hypothesis Results

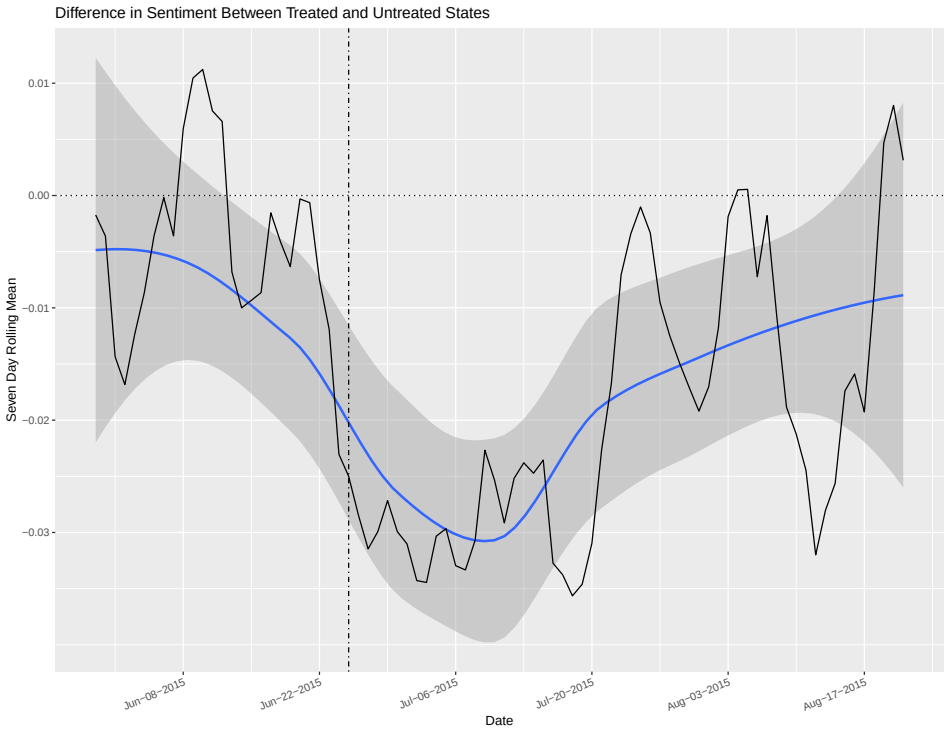
| | Positive Sentiment | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Unconstrained | Constrained | Unconstrained | Constrained |
| After | $0.049^{***}$ | $0.036^{***}$ | $-0.137^{***}$ | $-0.095^{***}$ |
| | $(0.011)$ | $(0.006)$ | $(0.023)$ | $(0.012)$ |
| Male | $-0.154^{***}$ | $-0.176^{***}$ | $-0.047^{*}$ | $0.031^{***}$ |
| | $(0.013)$ | $(0.004)$ | $(0.024)$ | $(0.010)$ |
| Male*After | $-0.025^{*}$ | | $0.093^{***}$ | |
| | $(0.014)$ | | $(0.026)$ | |
| Black | $-0.009$ | $-0.017$ | $0.027$ | $0.016$ |
| | $(0.038)$ | $(0.013)$ | $(0.074)$ | $(0.027)$ |
| Black*After | $-0.009$ | | $-0.013$ | |
| | $(0.041)$ | | $(0.079)$ | |
| Hispanic | $0.114^{***}$ | $0.170^{***}$ | $0.105^{*}$ | $0.117^{***}$ |
| | $(0.026)$ | $(0.009)$ | $(0.055)$ | $(0.022)$ |
| Hispanic*After | $0.063^{**}$ | | $0.015$ | |
| | $(0.028)$ | | $(0.060)$ | |
| Asian | $0.225^{***}$ | $0.224^{***}$ | $0.193^{**}$ | $0.201^{***}$ |
| | $(0.043)$ | $(0.014)$ | $(0.085)$ | $(0.034)$ |
| Asian*After | $-0.001$ | | $0.010$ | |
| | $(0.046)$ | | $(0.093)$ | |
| Other | $-0.045$ | $-0.130^{**}$ | $0.162$ | $-0.044$ |
| | $(0.216)$ | $(0.062)$ | $(0.506)$ | $(0.232)$ |
| Other*After | $-0.093$ | | $-0.266$ | |
| | $(0.226)$ | | $(0.570)$ | |
| GOP | | | $-0.508^{***}$ | $-0.553^{***}$ |
| | | | $(0.023)$ | $(0.009)$ |
| GOP*After | | | $-0.053^{**}$ | |
| | | | $(0.025)$ | |
| Constant | $0.973^{***}$ | $0.985^{***}$ | $0.903^{***}$ | $0.867^{***}$ |
| | $(0.011)$ | $(0.007)$ | $(0.021)$ | $(0.013)$ |
| Chi-Squared | $-$ | $8.733$ | $-$ | $16.416$ |
| Significant | $-$ | $0.044$ | $-$ | $0.004$ |
| N | $481,487$ | $481,487$ | $89,585$ | $89,585$ |
| Log Likelihood | $-224,589.100$ | $-224,593.500$ | $-51,406.950$ | $-51,415.160$ |

Table 2: Difference-In-Difference Analysis Results

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Positive Sentiment | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| After | $0.010^{***}$ | $-0.040^{***}$ | $0.007^{***}$ | $-0.003$ | $-0.027^{***}$ |
| | $(0.001)$ | $(0.002)$ | $(0.002)$ | $(0.003)$ | $(0.004)$ |
| Treated | $-0.003$ | $-0.002$ | $-0.009^{***}$ | $0.016^{***}$ | $-0.006$ |
| | $(0.002)$ | $(0.003)$ | $(0.003)$ | $(0.005)$ | $(0.008)$ |
| Treated*After | $-0.013^{***}$ | $-0.015^{***}$ | $-0.009^{***}$ | $-0.025^{***}$ | $-0.0004$ |
| | $(0.003)$ | $(0.003)$ | $(0.003)$ | $(0.006)$ | $(0.008)$ |
| GOP | | | | $-0.173^{***}$ | $-0.180^{***}$ |
| | | | | $(0.002)$ | $(0.003)$ |
| Constant | $0.847^{***}$ | $0.844^{***}$ | $0.875^{***}$ | $0.840^{***}$ | $0.858^{***}$ |
| | $(0.001)$ | $(0.001)$ | $(0.002)$ | $(0.003)$ | $(0.004)$ |
| Drop June 26? | No | Yes | No | No | No |
| Race and Gender | No | No | Yes | No | Yes |
| N | 1,028,151 | 673,057 | 481,487 | 184,042 | 89,585 |
| $R^2$ | 0.0004 | 0.002 | 0.004 | 0.040 | 0.046 |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

# Figures

Figure 1: Time Trends in Sentiment Across Treated and Untreated States

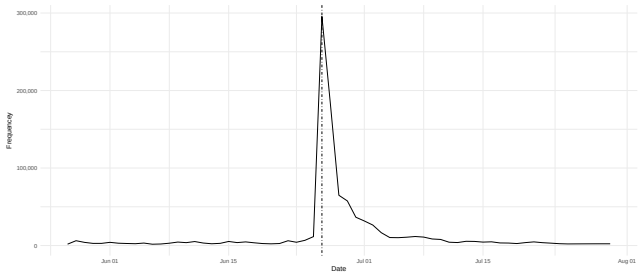


# Appendix A: Additional Descriptive Statistics

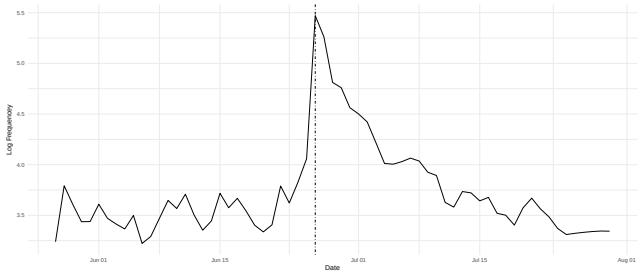
I collected the tweets analyzed in my project over a two-month time span, from May 27, 2015 to August 24, 2015. To obtain this data, I use a series of python scripts that continuously interacted with the Twitter API, using regular expressions to archive any tweet that contained one of the following issue words: **gay marriage, gay marriages, same-sex marriage, same-sex marriages, same sex marriage, same sex marriages, same-sex union, same-sex unions, same sex union, same sex unions, marriage equality, equal marriage**. During this time, I collected a total of 5,996,741 tweets. After filtering for location in the process described in the *Gathering Twitter Data* section above, I end up with 1,028,151 total tweets. In Figure A1, I plot the number of tweets I collected each day. The top half of Figure A1 plots the raw frequency of daily tweets, and it is immediately apparent that a very large number of tweets were sent on June 26, 2015, the day the Supreme Court announced their decision. This drops off quickly, although I collect a large number of tweets until early July. The bottom half of Figure A1 plots the logged frequencies in order to better visualize the entire time series.

Each state is represented in my dataset, with the number of tweets sent from each state enumerated in Table A1. One can also get a general sense of the distribution of users by looking at the heat map in Figure A2, which maps the number of tweets sent per capita using state populations recorded in the 2010 census.

Figure A1: Frequencies: May 27, 2015 to July 31, 2015




(a) Raw Frequency




(b) Log Frequency

Figure A2: Frequency of Tweets by State per Capita

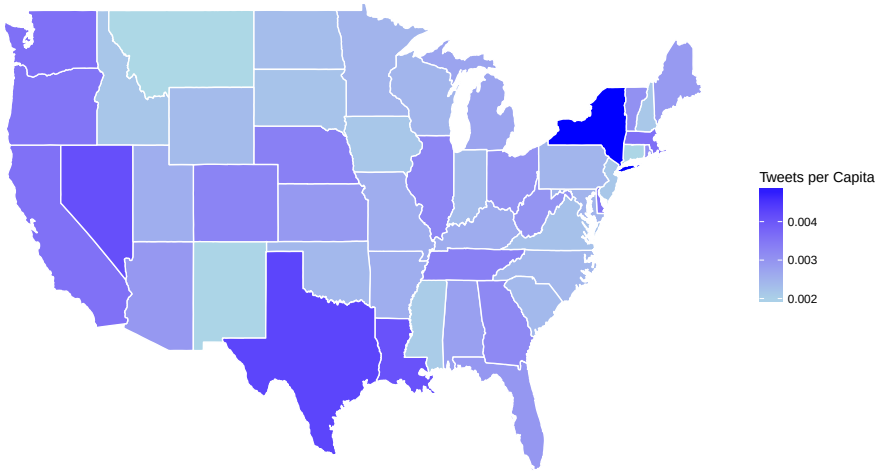


Table A1: Number of Tweets from each State

| State | Number of Tweets | State | Number of Users |
|---|---|---|---|
| California | 136,340 | Kentucky | 12,133 |
| Texas | 109,485 | Nevada | 11,425 |
| New York | 98,792 | South Carolina | 11,425 |
| Florida | 56,932 | Oklahoma | 9,184 |
| Illinois | 42,874 | Kansas | 86,25 |
| Ohio | 36,222 | Arkansas | 7,658 |
| Pennsylvania | 32,802 | Utah | 7,315 |
| Georgia | 31,259 | Connecticut | 7,005 |
| Washington D.C. | 29,888 | Iowa | 6,730 |
| Michigan | 27,755 | Mississippi | 6,239 |
| Washington | 25,295 | Nebraska | 6,203 |
| Massachusetts | 23,978 | West Virginia | 5,857 |
| North Carolina | 23,539 | New Mexico | 4,041 |
| Tennessee | 21,697 | Maine | 4,006 |
| New Jersey | 20,122 | Hawaii | 3,758 |
| Arizona | 19,372 | Idaho | 3,443 |
| Louisiana | 18,852 | Rhode Island | 3,235 |
| Virginia | 18,602 | Delaware | 3,189 |
| Colorado | 16,936 | New Hampshire | 2,909 |
| Missouri | 15,949 | Alaska | 2,428 |
| Indiana | 154,66 | Vermont | 2,001 |
| Maryland | 14,605 | Montana | 1,887 |
| Wisconsin | 14,391 | South Dakota | 1,851 |
| Alabama | 14,104 | North Dakota | 1,592 |
| Oregon | 14,008 | Wyoming | 1,351 |
| Minnesota | 13,389 | | |

# Appendix B: Preprocessing Text Data

Before running supervised training methods to estimate sentiment, I use several preprocessing scripts to manipulate and simplify the Twitter text data. First, I remove all textual information that does not inform the substance of the message, including punctuation, all forms of capitalization, and words that fail to contribute towards a sentence's meaning (such as "the, of, or").

Next, I tokenize the text, a process that splits "a string into its desired constituent parts" (Potts, 2011). My tokenizing strategy utilizes white-space to break apart a sentence into separate words. This transfers the content of a tweet into a list of individual words, ignoring the original order these words appear in the sentence. While the order of words in a sentence can absolutely contribute to the content of a message, treating each document as coming from a "bag-of-words" is a common (though at times contentious) assumption that is necessary to apply many machine learning methodologies (Grimmer & Stewart, 2013). In many situations, enough information can be gleaned from the choice of unique words to justify this assumption.

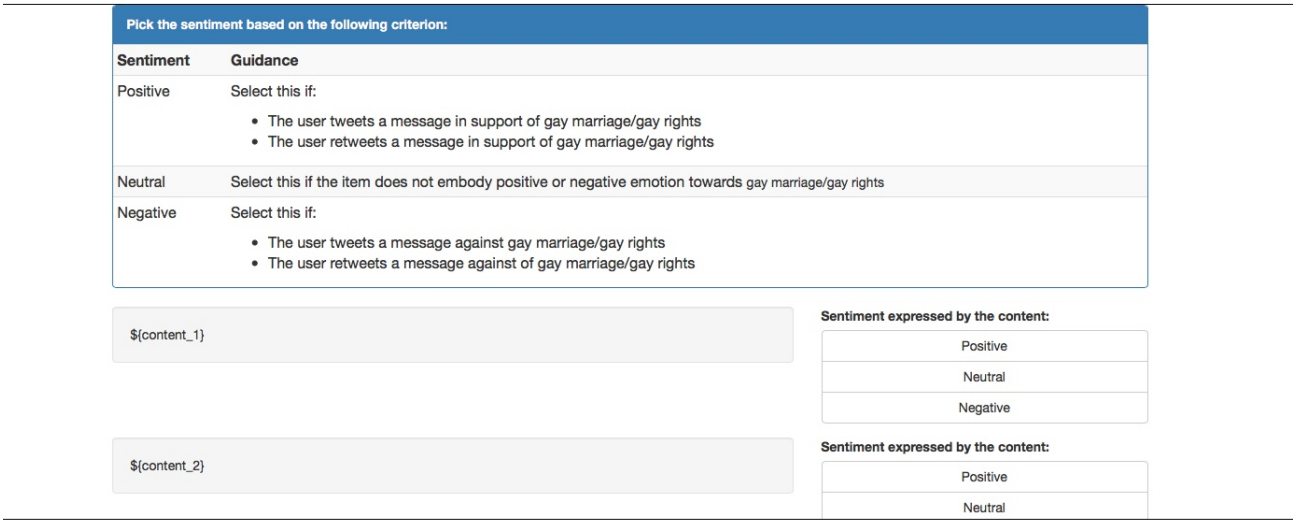
Finally, the entire dataset is transformed into a document-frequency matrix (DFM). A DFM is an $N \times J$ matrix, where $N$ is the number of documents (in this case, tweets) and $J$ is the number of unique features (in this case, individual words) found across all documents. Thus, if tweet $n$ contains two instances of word $j$, the $nj^{th}$ entry of the DFM is 2. With Twitter data, this represents a very sparse matrix, as the entire set of unique words $J$ across the entire dataset can be quite large, although an individual tweet being capped at 140-characters contains a small number of individual words (while Twitter eventually increased this cap to 280-characters, this occurred after my data collection period) Thus, rather than utilizing each of the $J$ unique features in the entire dataset, I analyze a subset of features based on how frequently the feature appears. This parameter can be tuned, but for the baseline analysis I kept a feature if it appeared at least three times throughout the dataset. In order to implement the preprocessing steps described above, this project utilized the `quanteda` R package (Benoit & Nulty, 2016). The `quanteda` package provides tools to organize and analyze string data in order to implement sentiment analysis methodologies.

# Appendix C: Validating the Supervised Scoring Method

In order to train a supervised classifier, I create a set of hand-annotated tweets using Mechanical Turk. In order to label the largest number of messages in the shortest amount of time, the set of annotated tweets corresponds with the top-4,000 most repeated messages in the dataset. In total, these 4,000 tweets represent 1,895,554 total messages, and thus consists of 31.60% of all collected tweets. After stripping these 4,000 messages of usernames, hyperlinks, and punctuation, there were 3,934 unique messages in the validation set.

In order to build this hand-annotated validation set, I utilized Amazon Mechanical Turk, a crowdsourcing platform that allows a researcher to pay individuals to complete small tasks. I created a set of tasks that required Mechanical Turk users to score the sentiment of ten tweets in my validation set. I present a screen shot of the task in Figure A3.

Figure A3: Sample Mechanical Turk Task



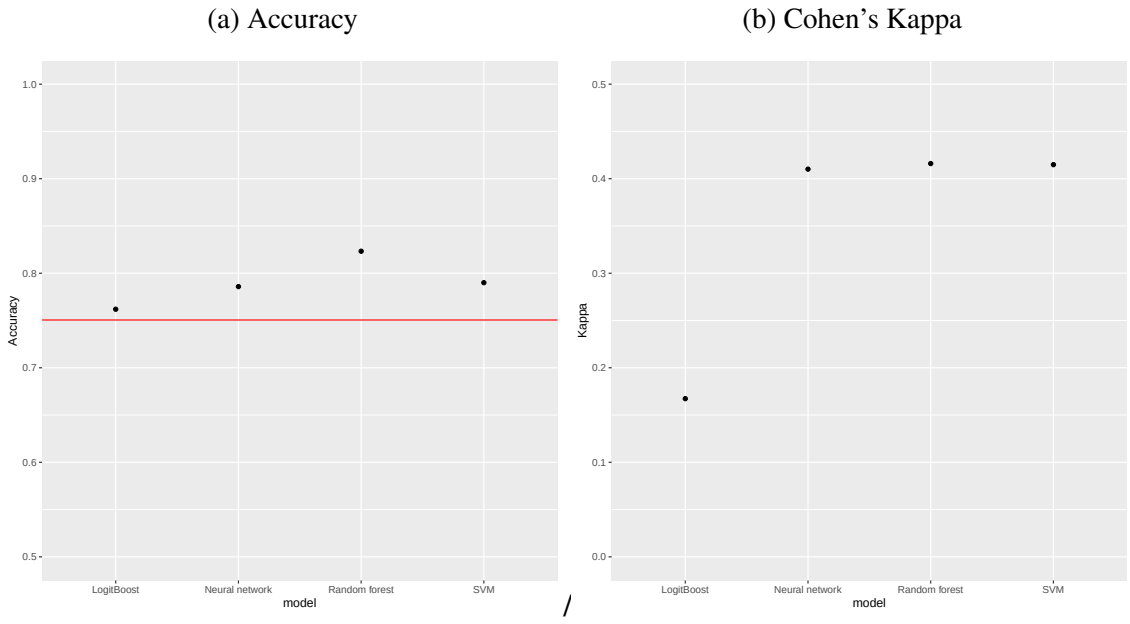

Each task was performed by three separate Mechanical Turk users in order to get a sentiment score as close to the ground truth as possible. To create a final score for each of the 3,934 unique messages, I took the majority score across the three annotations. In total, the Mechanical Turkers labeled 626 messages negative, 1,778 positive, and 1,333 neutral. Only 197 messages did not have

a majority category. In the body of the paper, I focus on using the 626 negative and 1,778 positive messages to train a binary classifier. In Appendix E, I use the neutral messages to build a three-way classifier.

In order to train a classifier, I split 10% of the training data to use a test set. This test set of tweets is not used to train the final model, and thus allows me to evaluate the performance of the model on new data.

I start by testing a number of different classifiers, including Support Vector Machines, logit-boost, neural networks, and random forest.[1] For each model, I run a 10-cross fold validation to train the classifier before evaluating performance on the left out set. The accuracy (compared with a no-information red in red) and kappa coefficient of the best performing models in each category are found in Figure A4. I find that random forest leads to the highest accuracy without sacrificing inter-rater reliability.

Figure A4: Comparing Classifiers




On choosing an overall model classification, I tune the hyper-parameters of the random forest model. Repeating 10-cross fold validation 10 times per hyper-parameter, I test which minimum

[1] I train and evaluate all classifiers with the `caret` package (Kuhn, 2008)

node size leads to the best cross validation accuracy.

For the test set, my final model accurately predicts 81.74% of the data, with 97.78% Sensitivity and 34.43% Specificity.

To better diagnose the model, I present the receiver operating characteristic (ROC) curve (with area under curve reported) in Figure A5, and the test-set confusion matrix in Table A2.

Figure A5: ROC curve

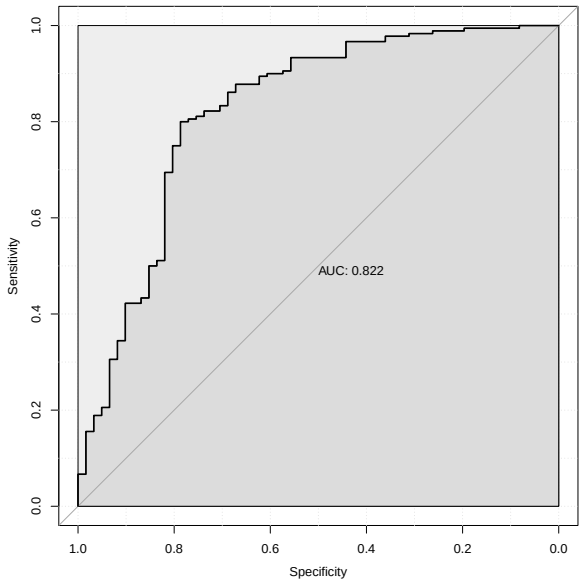



Table A2: Error Matrix:
Test Set Predictions

| | **Predicted Negative** | **Predicted Positive** | **Total** |
|---|---|---|---|
| **True Negative** | 21 | 4 | 25 |
| **True Positive** | 40 | 176 | 216 |
| **Total** | 61 | 180 | |

The error matrix reveals that one issue my classifier exhibits is over-predicting the positive class. While part of this issue may stem from the fact I have an unbalanced training set, visually

inspecting may of the false-positive tweets reveals that many misclassified messages are highly sarcastic in tone. While this is easy for a human reader to recognize, sarcasm is very difficult to detect in sentiment scoring algorithms.[2] Overall, this reveals that my classifier is more likely to falsely classify a negative tweet as positive, biasing all my scores upwards. Thus, as my core finding is finding a more *negative* reaction in states with a law change, this upward bias is likely attenuating my findings. Thus, this bias should not hurt the causal interpretation of my core results.

# Appendix D: Neutral Tweets

An issue potentially biasing my results is the presence of a third sentiment category: neutral messages. While theoretically possible to build a third training set of *neutral* tweets and training a three-way classifier, binary classifiers tend to lead to more accurate labels. However, in a robustness check, I retrain the classifier using the neutral labels collected on Mechanical Turk. In total, this training set consists of 626 negative, 1,778 positive, and 1,333 neutral tweets.

I train this model with the same procedure described in an earlier section: leaving out 10% of the data as a test set, and doing 10-fold validation across the training set to tune over the hyperparameters in the random forest model. The confusion matrix for the best performing model on the left out test set is found in Table A3. In total, the model has an accuracy of 61.23% against a 47.06% no information rate, and Cohen's kappa coefficient 0.35.

Applying this three-way classifier to my analysis, I rerun my main model specification including with neutral labels. I score neutral messages as 0.5, in addition to scoring negative messages 0 and positive messages 1. I present the results of this robustness check in Table A4

In Table A4, I note that, across each model specification, the **Treated**$\times$**After** coefficient remains negative and statistically significant. In fact, the model including neutral labels more robustly demonstrates my core results, finding a small level of near statistical significance in model five (a null result in the binary model). This seems to provide additional evidence that the inclusion of neutral tweets biases my core results upwards, allowing me to better interpret the core results in

[2] See (Maynard & Greenwood, 2014) as an example of one attempt to address sarcasm detection in tweets.

Table A3: Multiclass Model Error Matrix:
Test Set Predictions

| | **Predicted Negative** | **Predicted Neutral** | **Predicted Positive** | **Total** |
|---|---|---|---|---|
| **True Negative** | 15 | 6 | 2 | 23 |
| **True Neutral** | 31 | 89 | 49 | 169 |
| **True Positive** | 18 | 39 | 125 | 182 |
| **Total** | 61 | 134 | 176 | |

Table A4: Difference-in-Difference Results: Three-Way Classifier

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Positive Sentiment | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| After | $0.064^{***}$ | $0.016^{***}$ | $0.055^{***}$ | $0.030^{***}$ | $0.023^{***}$ |
| | (0.001) | (0.001) | (0.002) | (0.002) | (0.003) |
| Treated | 0.001 | 0.002 | $-0.0004$ | $0.007^{*}$ | 0.0002 |
| | (0.002) | (0.002) | (0.003) | (0.004) | (0.006) |
| Treated*After | $-0.019^{***}$ | $-0.026^{***}$ | $-0.018^{***}$ | $-0.016^{***}$ | $-0.010^{*}$ |
| | (0.002) | (0.002) | (0.003) | (0.004) | (0.006) |
| GOP | | | | $-0.164^{***}$ | $-0.155^{***}$ |
| | | | | (0.001) | (0.002) |
| Constant | $0.649^{***}$ | $0.648^{***}$ | $0.669^{***}$ | $0.643^{***}$ | $0.646^{***}$ |
| | (0.001) | (0.001) | (0.002) | (0.002) | (0.003) |
| Drop June 26? | No | Yes | No | No | No |
| Race and Gender | No | No | Yes | No | Yes |
| N | 1,076,512 | 673,792 | 506,274 | 191,980 | 93,681 |
| $R^2$ | 0.004 | 0.001 | 0.011 | 0.062 | 0.060 |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

a causal manner.

# Appendix E: Border State Analysis

As the parallel trend assumption in the difference-in-difference estimator posits that the untreated group is a good counter-factual to the treatment group, a potential criticism of my work is that I do not restrict the group of untreated states. That is, I analyze data from all fifty states, when perhaps states like California and New York do not make good counterfactuals to the states in the treatment group.

While a matching methodology represents the most rigorous way to find valid counterfactuals for users in my treated set, I do not have a rich enough set of independent variables to allow for an accurate matching procedure. However, it is possible to use the geography of the treated states to find a set of users that might represent a more valid counterfactual. Thus, I re-run my analysis with a smaller set of untreated states, restricting the untreated group to only those states that share a border with one or more treated states.[3] By restricting the untreated states in this way, I am more likely to select states with similar demographic characteristics, allowing me to further test and validate my results. The result of this robustness check is found in Table A5, which replicates the model specifications in the body of the paper.

In models one and two, the baseline models, I find a negative and statistically significant **Treated**$\times$**After** coefficient. Thus, even when restricting the untreated group to smaller set of states more likely to share characteristics with the treated set, I continue to find evidence of a causal impact. I also find this impact in model three, where I include demographic information. In models four and five, where I include the partisan labels, I find a null result. This is partially due to the upward bias of the binary classifier described in the previous appendix; with a higher probability in coding neutral messages as positive, the results are biased upwards, away from my hypothesis.

To test the impact of neutral messages on the border states, I rerun the analysis with the three-

[3]The bordering states include: Oklahoma, Kansas, New Mexico, Colorado, Wyoming, Montana, Minnesota, Iowa, Wisconsin, Illinois, Indiana, Alabama, Florida, South Carolina, North Carolina, Virginia, West Virginia, Pennsylvania.

Table A5: Border States with Binary Classifier

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Positive Sentiment | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| After | $0.015^{***}$ | $-0.053^{***}$ | $0.023^{***}$ | $-0.033^{***}$ | $-0.037^{***}$ |
| | $(0.002)$ | $(0.003)$ | $(0.003)$ | $(0.005)$ | $(0.007)$ |
| Treated | $0.007^{**}$ | $0.008^{**}$ | $0.015^{***}$ | $-0.009$ | $0.0004$ |
| | $(0.003)$ | $(0.003)$ | $(0.004)$ | $(0.007)$ | $(0.010)$ |
| Treated*After | $-0.013^{***}$ | $-0.015^{***}$ | $-0.022^{***}$ | $0.00001$ | $0.005$ |
| | $(0.003)$ | $(0.004)$ | $(0.005)$ | $(0.007)$ | $(0.011)$ |
| GOP | | | | $-0.212^{***}$ | $-0.205^{***}$ |
| | | | | $(0.003)$ | $(0.004)$ |
| Constant | $0.800^{***}$ | $0.797^{***}$ | $0.813^{***}$ | $0.833^{***}$ | $0.807^{***}$ |
| | $(0.002)$ | $(0.002)$ | $(0.003)$ | $(0.005)$ | $(0.007)$ |
| Drop June 26? | No | Yes | No | No | No |
| Race and Gender | No | No | Yes | No | Yes |
| N | 580,896 | 362,785 | 272,415 | 102,474 | 49,974 |
| $R^2$ | 0.0001 | 0.003 | 0.005 | 0.055 | 0.052 |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

way classifier described in Appendix D: *Neutral Tweets*. In this model, I code the dependent variable as 0 for negative tweets, 0.5 for neutral tweets, and 1 for positive tweets. I present the results in Table A6.[18]

Table A6: Border States with Three-Way Classifier

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Positive Sentiment | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| After | $0.065^{***}$ | $0.015^{***}$ | $0.066^{***}$ | $0.026^{***}$ | $0.031^{***}$ |
| | (0.002) | (0.002) | (0.003) | (0.003) | (0.005) |
| Treated | $0.014^{***}$ | $0.015^{***}$ | $0.024^{***}$ | 0.008 | $0.024^{***}$ |
| | (0.002) | (0.003) | (0.003) | (0.005) | (0.006) |
| Treated*After | $-0.021^{***}$ | $-0.025^{***}$ | $-0.028^{***}$ | $-0.011^{**}$ | $-0.019^{***}$ |
| | (0.003) | (0.003) | (0.004) | (0.005) | (0.007) |
| Republican | | | | $-0.175^{***}$ | $-0.160^{***}$ |
| | | | | (0.002) | (0.003) |
| Constant | $0.637^{***}$ | $0.635^{***}$ | $0.647^{***}$ | $0.647^{***}$ | $0.621^{***}$ |
| | (0.002) | (0.002) | (0.003) | (0.003) | (0.005) |
| Drop June 26? | No | Yes | No | No | No |
| Race and Gender | No | No | Yes | No | Yes |
| N | 607,695 | 378,437 | 286,118 | 106,761 | 52,115 |
| $R^2$ | 0.003 | 0.0003 | 0.012 | 0.072 | 0.064 |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Here, I find that all the results are extremely similar to Table A5, but with a negative and significant **Treated**×**After** coefficient in models 4 and 5. This robustness check helps confirm the results in the main section of my paper.

# Appendix F: Checking for Bot Accounts

Detecting 'bot' accounts is the subject of many machine learning papers, with researchers focusing on different techniques to determine whether messages are sent by humans or automated programs (e.g. Wang, 2010; Jajodia et al., 2012; Ferrara, Varol, Davis, Menczer, & Flammini, 2016). Given the discussions in the wake of the 2016 U.S. election regarding automated systems disseminating "fake news" on social media platforms, it is important to consider whether or not my dataset is filled with 'bot' accounts biasing my results.

To get a sense of how many likely bot accounts are present in my dataset, I pull a sample of 30,000 random users. To figure out how likely these 30,000 users are 'bot' accounts, I utilize the `Botometer` publicly available API.[4] The `Botometer` API interacts with the Twitter API, pulling over one thousand features from the user's Twitter profile to compare against a collection of 15,000 manually verified bot accounts and 16,000 verified human accounts (Varol, Ferrara, Davis, Menczer, & Flammini, 2017). The classifier then runs an ensemble method using random forests, AdaBoost, logistic regression, and decision trees to determine the likelihood a given user is human or a 'bot.' The classifier outputs a likelihood from zero to one; the closer the bot score is to one, the more likely the account is run by an automated program. I present the distribution of classification scores from 30,000 randomly selected users in Figure A6.

Figure A6 demonstrates that the majority of users are likely human, with a mean bot score of 0.29 with a standard deviation of 0.14 across the sample. Only a small number of users are likely bots, with only 9.2% of users with a bot score greater than 0.5 and 1.3% of users with a bot score greater than 0.75. While important to note `Botometer` represents only one approach to detecting bots, this preliminary analysis shows little evidence that bots drive my results.

[4] https://botometer.iuni.iu.edu

Figure A6: Histogram of Twitter Bot Likelihood