Measuring Anti-LGBTQ+ Language on Twitter

Joe Marlo, George Perrett, Bilal Waheed

December 2020

Abstract

The divisive and polarizing rhetoric in the 2016 presidential election sparked concern over popularizing hateful sentiments towards marginalized populations on Twitter. In this paper, we focus on the LGBTQ+ community and examine ~100 million tweets for the presence of hate speech targeted towards LGBTQ+ Americans as a result of key political and social events related to the LGBTQ+ community. Dictionary-based methods refined by logistic regression, Naive Bayes, and Recurrent Neural Network (RNN) machine learning classifiers were used to identify hate speech. We found no conclusive evidence of changes in prevalence or incidence of hate speech around key events. While some events saw brief upticks in prevalence, overall levels of hate speech remained stable. Our analysis finds exploratory evidence of decreases in incidence of anti-LGBTQ+ hate speech ($p < 0.001$) over time coinciding with a Twitter policy change allowing users to directly report abuse.

Introduction

Hate speech on social media has been identified as a major problem, yet little is known about the prevalence and patterns of hate speech on social media sites. In a 2019 paper, Siegel and colleagues investigated the prevalence of racially motivated hate speech on Twitter in the months before, during and after the 2016 presidential election. We aim to replicate these research methods to investigate LGBTQ+ directed hate speech.

Motivation

The motivation for this research stems from the overarching question of whether or not people of the LGBTQ+ community feel safe living in the U.S. The result of the highly polarized and divisive 2016 presidential campaign led to concern that hateful language and attitudes were being sowed into the general public's mindset.

Anecdotal evidence from media coverage portrayed a narrative that Twitter and other social media platforms were hotbeds for political discord and a vector of hateful and divisive rhetoric. Due to president Trump's active presence on Twitter, minority populations felt increasingly unwelcome and threatened by the incumbent presidential administration.

In particular, the actions of the Trump administration led to fear of setbacks in the fight for LGBTQ+ rights and social equality. In its annual "Accelerating Acceptance" study in 2018, the U.S. based NGO GLAAD ("Gay & Lesbian Alliance Against Defamation, now formally "GLAAD" as of 2013) found diminishing levels of acceptance of LGBTQ+ people in society following the 2016 election, especially among younger people.

The purpose of this paper is to focus on the LGBTQ+ community and assess whether or not major political and social events during the Trump administration led to a rise in hateful language targeted towards them. We look to Twitter to see if there is any evidence at scale, and if there is any relationship between political events, social media, and targeted hate speech.

Research questions

1. Does the prevalence of anti-LGBTQ+ language on Twitter in the United States change with respect to political and social events?
    - Windsor v. U.S. Case - June 2013
    - Legalization of same-sex marriage - June 2015
    - Pulse nightclub shooting - June 2016
    - Election Day 2016 - November 2016
    - Inauguration Day - January 2017
    - Transgender military ban - July 2017

2. Is there a detectable baseline level of anti-LGBTQ+ language on Twitter?

Data collection and sampling methodology

The goal was to randomly sample the US population of Twitter users to obtain a representative sample of users and their timelines. We believed approximately 250,000 accounts would be necessary to obtain enough tweets that contain hate speech within the intervals we are interested in studying. The final dataset consists of approximately 100 million tweets spanning 160 thousand accounts from 2006 to 2020. Importantly, the tweets are captured by first identifying users and then collecting their tweet history (i.e. their timeline) rather than collecting tweets regardless of user.

Bias

A simple random sample is not possible due to technical limitations so our sampling strategy is quasi-random with a known bias. The frequency of collected tweets plateaus in 2013 and then starts increasing in 2020 (Figure 1) rather than monotonically increasing over time with Twitter popularity. The associated accounts appear to skew towards 2012-2014 account opening dates (Figure 2, derived from the first known tweet) which may explain the plateau. The latter increase is most likely due to a limitation of being able to collect only the latest 3,240 tweets from a given user.

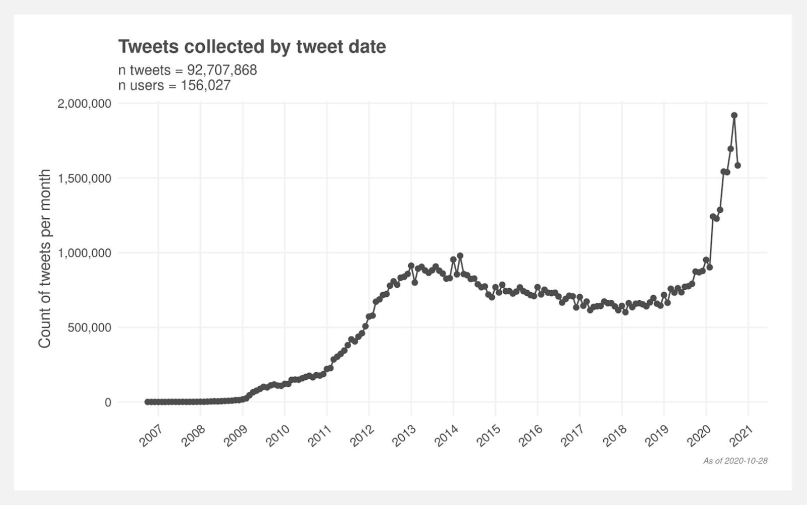

Figure 1: Tweets collected by tweet date

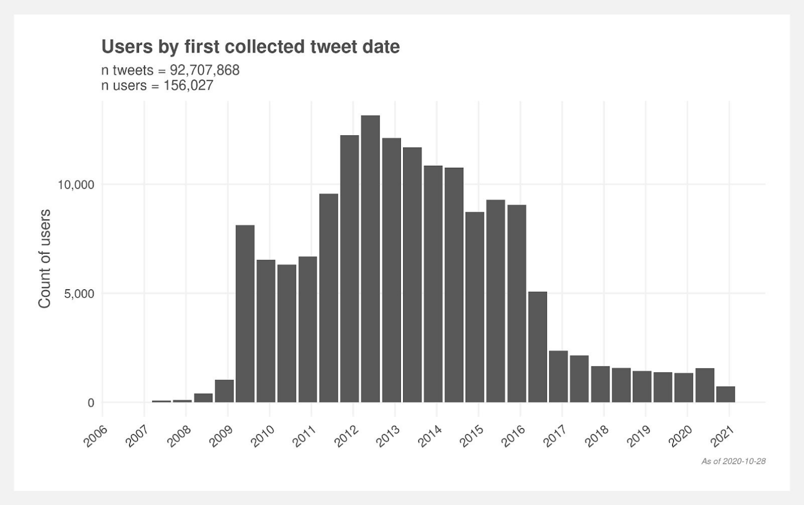

Figure 2: Twitter users by first collected by tweet date

Twitter IDs

The sample was achieved using the Twitter API to sample users based on their unique numeric ID. We first generated random numbers and tested the Twitter API to see if the user ID existed. Twitter generates the user IDs using a quasi-random sequentially-based method to produce a 64-bit integer (King 2010). This results in a number between 1 and 9 quintillion. This is not a computationally feasible sample space to sample from. However, IDs were previously generated using a method that produced much smaller numbers so the majority of IDs are in a feasible sample space. To determine our sample of IDs, we first trialled a series of numbers against the Twitter API to ascertain the possible distribution of the sample space, and then deduced that generating random numbers between 1 and 10 billion provides a balance between representativeness and computation time. Approximately 300 million random integers were then generated and checked against the Twitter API to verify if it is a valid ID and if the user's self-described open-text location field indicates a United States location. Location was verified by checking if the field mentions a full or abbreviated US state, the full name of one of the approximately 1,000 most populated US cities, or the country name.

Collection

The sampling was further hampered by the Twitter API restrictions which limits API calls to 900 per 15 minutes. A Python script executed on a Raspberry Pi microcomputer was used to continuously make API calls for approximately 25 days and store the resulting data in a SQLite database.

Identifying hate speech

We first establish an operational definition of hate speech. There is no universally accepted definition of hate speech, as academics and policymakers vary in their interpretations of what constitutes hate. For the purposes of this research, we focus on defining hate speech with a broad interpretation, in order to cover a variety of messages and expressions that target and justify hate towards a specific group of people (in this case, the LGBTQ+ population).

Prior research on this topic operationalized the definition of online hate speech by referring to definitions used by hateful conduct policies on social media platforms like Twitter and Facebook. By taking these into consideration, we define hate speech as *any language that is used to express, motivate, and justify hatred towards a person or group of people based on their perceived or actual identities, or is intended to offend, humiliate, or insult the person and/or members of the group*. (Davidson 2017)

This definition minimizes the inclusions of messages and tweets that express pride in one's own identity, endorsements of other hate-affiliated groups, and other ambiguities, to focus on targeted messages disparaging others. This also helps differentiate hate language from general offensive language. However, we examine the latter in our final analysis, which is discussed formally in our analysis section.

Hate speech dictionary

With this framework, the next task was to establish a working dictionary of terms and phrases that are associated with LGBTQ+ hate speech. We consulted hatebase.org which is an active database containing a lexicon of terms identified by internet users as hate speech.

The dictionary approach collects a range of tweets containing anti-LGBTQ+ language but also flags many tweets that do not include true instances of anti-LGBTQ+ language. Table 1 shows selected examples from the initial dictionary-based screener.

Tweet	**Description**
"rt @bhand_engineer: @zakirism kuch palo ke liye apne marg se bhatak gaya tha prabhu. aapko vishwaasghat dena humaara maqsat nahi tha. innoc…"	Tweet not in english
"can't believe i've been gay for 23 years and tomorrow is going to be my first time going to pride"	Non-negative LGBTQ+ Tweets
"when you wanna go out but all your friends are gay af"	More ambiguous case
"rt @chefpolohoe: u gay af for lettin dat shit buss all in yo mouth like dat https://t.co/xtlp6s8dri"	Explicit hate speech.

Table 1: Selected hate speech examples using dictionary-based approach.

Classifying anti LGBTQ+ language

A two step process was used to identify sampled tweets that use LGBTQ+ slurs. The first step involved flaggiging a broad set of candidate tweets that are possible instances. Regex pattern matching was used to flag tweets that contained language that resembled the pre-selected terms in the lexicon described above. Of the 92,707,868 collected tweets 167,724 (0.18%) were identified as potential incidents of LGBTQ+ directed slurs. This approach offers a crude approximation of instances of anti-LGBTQ+ speech. Notably, this approach is not able to separate tweets that explicitly condemn the use anti-LGBTQ+ language true incidents of slurs.

Several machine learning models were fit to disentangle true instances of anti-LGBTQ+ tweets from false positives identified by the regex. Prior to model fitting a 6,000 tweets were randomly sampled from the 167,724 tweets flagged in the first step. These 6,000 were coded by a team of three researchers. The coding scheme identified each of the 6,000 tweets as either a true instance of anti-LGBTQ+ language or as a false positive. These labels were then used as training data for logistic regression word embeddings, Naive Bayes and RNN machine learning models. Of the 6,000 labeled tweets, 10% were removed and saved to assess out of sample accuracy while the remaining were used to train each of the candidate models outlined below.

Logistic regression with word embeddings

Text data can be modeled within a logistic regression context by using a bag of words approach. Under this approach, a document term matrix is created from the training set and the words that make up each tweet are dummy coded for all variables. Logistic regression then determines the probability of a given tweet by fitting probabilities between words and respective classifications. Shown below, logistic regression transforms probabilities into linear predictors. In the case of text classification, each word in the bag of words matrix provides a linear coefficient representing its likelihood association with anti-LGTQ+ coding, the sum of all words included in a tweet can be converted to probability and used to classify the likelihood that a given tweet is epressignanti-LGBTQ+ sentiment.

$$\boldsymbol{\ell} = log_b \frac{p}{1-p} = \beta_0 + \beta_X$$

Naive Bayes

Naive Bayes is a classic method for text classification problems and has been successfully used to separate spam from real email (Dhinakaran, Nagamalai & Lee, 2009) as well as identify hate speech in political science research(Seigel et al., 2019). Naive Bayes works by creating a term matrix of words from training data. The probability of each word coinciding with training labels is calculated. Predictions are the multivariate probabilities of all words in a given sample of text. In the context of this problem, probabilities of a given word occurring in true instances of Anti-LGBTQ+ speech are compared to the probability of that word occurring in a false positive.

For a given tweet the multivariate probabilities of each world are combined to predict if the tweet is more likely to be a true case of anti-LGTQ+ speech or a false positive. Naive Bayes is referred to as ‘naive’ because it implicitly assumes that all columns (in this context, words) are independent from one another. This strong assumption allows the necessary calculations to be computationally possible, however, this assumption is never actually met. Despite the inter-column dependencies and the known violation of the independence assumption, Naive Bayes has repeatedly performed well on text classification problems. After making this assumption of joint independence, the equation below can be used to calculate probabilities of each class. C represents the classification (1 for anti-LGBTQ+ language, 0 for non-anti-LGBTQ+ language) while x corresponds to each word in the bag of work feature set.

$$p(C_k|x_1, ..., x_n) \propto p(Ck) \prod_{i=1}^{n} p(x_i|C_k)$$

Recurrent Neural Network (RNN)

Recently, Neural Networks and deep learning have become the leading tool for text classification purposes. While there are many different types of neural network architectures, the sequenced based nature of RNNs makes them particularly applicable for short text classifications(Gelron, 2019). Hyperparameters, the number to layers, length of sequences, and number of weights were selected through cross validation on a validation set (different than the 10% held out for out of sample testing).

The RNN was trained to maximize AUC on the out of sample validation set. In the context of twitter data, RNNs work by using each tweet as the unit of data. The advantage of the RNN model is its ability to incorporate ordering and context of prior words. In effect, RNNs are able to test all possible combinations of N-grams, an approach that would be infeasible for a bag of words based approach. RNNs are able to identify relationships by looping over a sequence of text. Shown in the RNN diagram below, a given tweet enters the model at input layer x_i, the sequence of text is then looped through layers 2 and 3. This looping allows the model to incorporate information from across the sequence in a simultaneous process. Backpropagation is used to determine the weights of each node, both within and outside of the recurrent layers, and binary predictions are released using a softmax transformation at output node y.

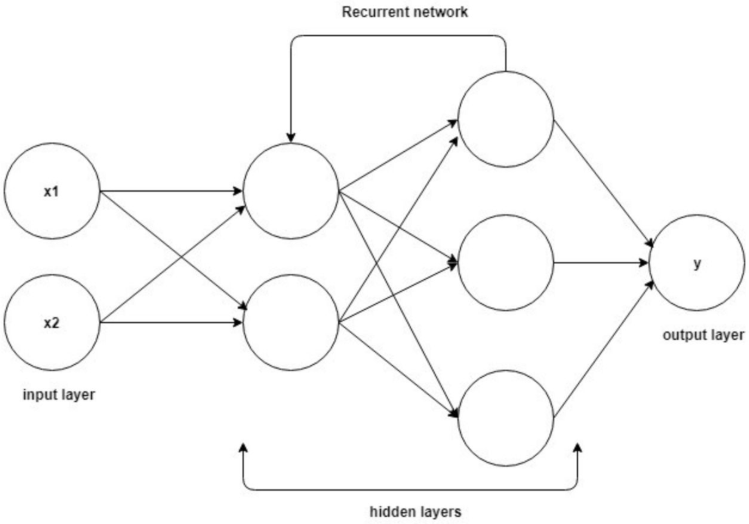

Figure 3: Diagram of RNN

Comparing models

AUC, precision, and recall of all four methods are shown in the table below.

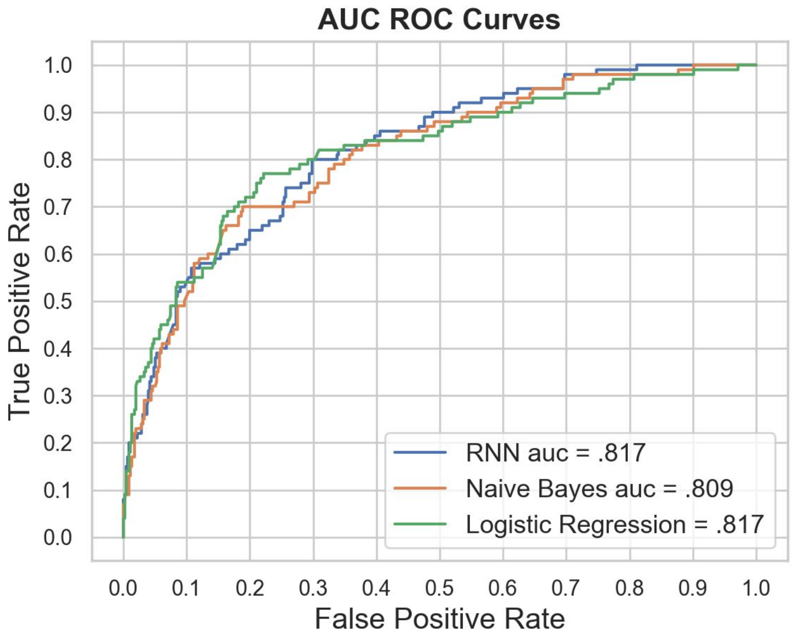

Figure 4: Comparing model AUC

Model	Precision	Recall
RNN	0.53	0.56
Naive Bayes	0.66	0.23
Logistic Regression	0.66	0.39

Table 2: Model diagnostics

Tweet	Description	Classification
“rt @bhand_engineer: @zakirism kuch palo ke liye apne marg se bhatak gaya tha prabhu. aapko vishwaasghat dena humaara maqsat nahi tha. innoc…”	Tweet not in english	Not hate speech
“can’t believe i’ve been gay for 23 years and tomorrow is going to be my first time going to pride”	Non-negative LGBTQ+ Tweets	Not hate speech
“when you wanna go out but all your friends are gay af”	More ambiguous case	Hate speech
“rt @chefpolohoe: u gay af for lettin dat shit buss all in yo mouth like dat https://t.co/xtlp6s8dri”	Explicit hate speech.	Hate speech

Table 3: Model classification results of selected hate speech examples

Results

Of the three candidate models, we selected the RNN due to its balance in recall and precision. Moreover, as more labeled training tweets are obtained, the predictive power of the RNN will increase. The RNN model identified 32,554 tweets or 0.035% of the sample as containing anti-LGBTQ+ language. This is the equivalent of 35 in 100,000 tweets. This prevalence rate varies over time, generally increasing in the early years of Twitter and then decreasing post 2013. Figure 5 shows the change in prevalence between 2008 and 2020.

There is a significant amount of noise in the early years which we believe is driven by the relatively small sample size. There’s a clear inflection point in the rate in mid 2013. The rate increases up until then and then monotonically decreases reaching apparent stability around 2017. Visually, there appears to be local upticks in the rate in mid 2015, mid 2019, and mid 2020. The vertical dashed lines in figure 5 represent the six key dates of interest.

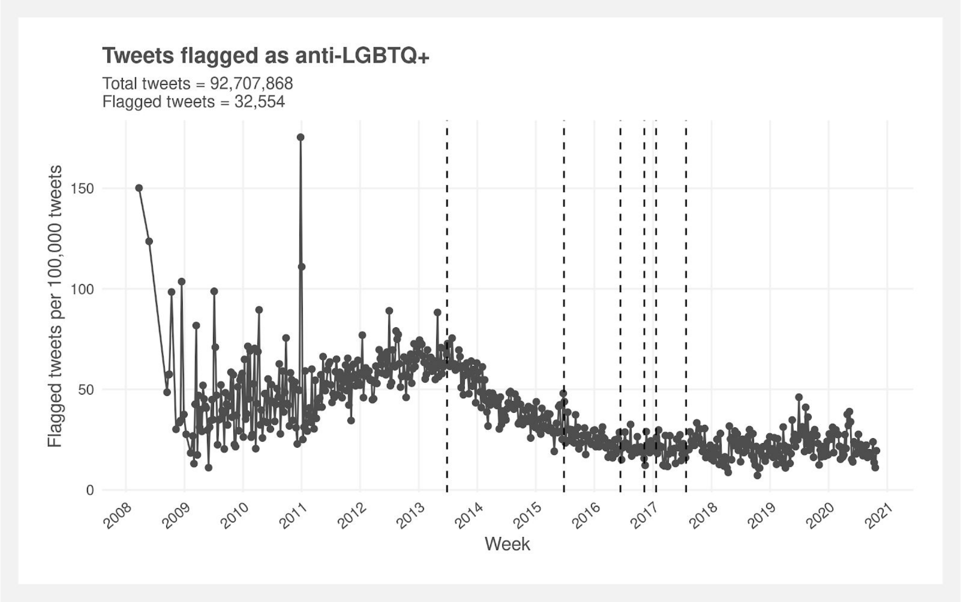

Figure 5: Proportion of tweets identified as anti-LGBTQ+

Key dates

The inflection point in mid 2013 is the most visually evident trend. Two other dates also show local changes in the prevalence rate: mid 2015 (legalization of same-sex marriage) and mid 2017 (transgender ban). Figure 6 shows the prevalence rate centered around the six key dates with a two-year bandwidth.

The mid 2013 inflection corresponds to the ruling of Windsor vs the United States. The case established that Section 3 of the Defense of Marriage Act (DOMA) is unconstitutional. This resulted in the prevention of federal discrimination against gay and lesbian couples for the determination of federal benefits and protections. It notably occurs two years before the repeal of DOMA and full legalization of same-sex marriage (mid 2015).

The legalization of same-sex marriage resulted from the landmark case of *Obergerfell v. Hodges* in 2015. The U.S. Supreme Court ruled state bans on same-sex marriage as unconstitutional, making same-sex marriage legal throughout America.

In July, 2017, President Trump announced via Twitter that the U.S would no longer accept or allow transgender individuals to openly serve in any capacity in military service, citing an increased burden on spending from prescribed drugs and medications.

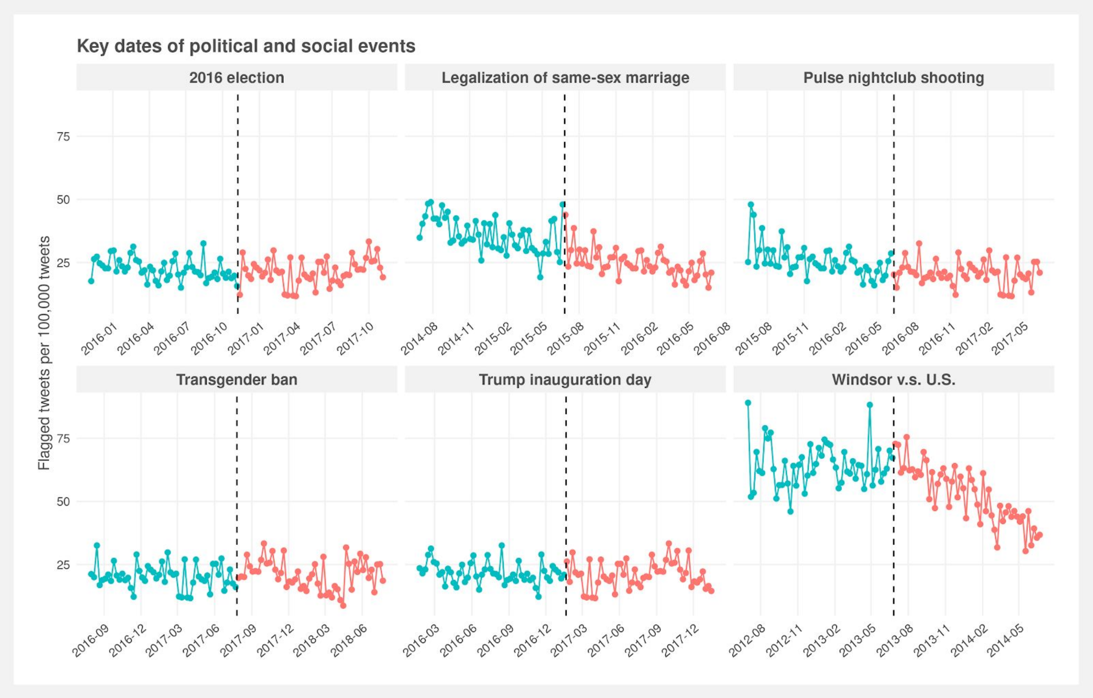

Figure 6: Proportion of tweets identified as anti-LGBTQ+ for key dates

Analysis

Changes in anti-LGBTQ+ language after the identified time-points can be quantified through changes in prevalence and changes in incidence. Prevalence reflects shifts immediately following one of the identified events, while incidence quantifies changes in rate of anti-LGBTQ+ tweets over time.

We used a regression discontinuity design (RDD) to measure changes in the prevalence of anti-LGBTQ+ tweets immediately after each of the six time points. Shown below, an RDD model fits local regressions to each side of an established cut point. For each of the six models, the date of the event serves as the established cutpoint, *T* refers to a binary variable representing data points either before or after the event of interest while *r* represents how far a respective data point is from the established cut point, prior to fitting each model *r* is centered such that β*Ti* represented the effect.

$$Y = \alpha + \beta_0 Ti + \beta_1 * r_i + \epsilon_i$$

Each of six RDD models tests for difference immediately prior to and after the identified events, however, this approach can not test changes in the rate of anti-LGBTQ+ tweets. To test changes in incidence we used an interrupted time series design. Shown below, the interrupted time series design includes variable βtime*event tests is the rate of anti-LGBTQ+ tweets changes following the identified event.

$$Y = \alpha + \beta_{Ti} + \beta_{event} + \beta_{Ti} * \beta_{event} + \epsilon_i$$

Both of these models require specifying a bandwidth of data points to include in the analyses. We used the R package rdrobust (Calonico et al., 2020) to select bandwidth via cross validation. The relevant coefficients testing incidence and prevalence for each of the six events are shown in the table 4. The Bonferroni adjustment was applied to all p values to control for multiple testing and maintain a significance threshold of 0.05.

	Incidence			**Prevalence**		
Event	β	p value	Adjusted p value	β	p value	Adjusted p value
2016 presidential election	0.183	0.002	0.028	0.193	0.923	1.000
Legalization of same-sex marriage	-0.032	0.680	1.000	-2.625	0.492	1.000
Pulse nightclub shooting	0.205	0.000	0.006	1.878	0.293	0.985
Transgender ban	-0.064	0.322	0.991	5.401	0.002	0.023
Trump inauguration day	0.157	0.023	0.248	-0.990	0.637	1.000
Windsor vs. US	-0.638	0.000	0.000	3.369	0.279	0.980

Table 4: Interrupted time series results

Shown in Table 4, we find evidence that the prevalence of anti-LGBTQ+ tweets increased immediately after the announcement of the transgender military service ban (β=5.401, p=0.023). There was no evidence that the prevalence of anti-LGBTQ+ tweets increased immediately after the 2016 election, the legalization of same sex marriage, the inauguration of President Trump or the decision of Windsor vs. US (p>0.05). We found evidence for an increase in incidence following the Pulse nightclub shooting (β=0.205, p<0.006) and the 2016 presidential election (β=0.183, p=0.028. Contradictory to expectations, we found evidence of a sharp increase in the incidence of anti-LGBTQ+ hate speech following the decision of Windsor vs. US (β=-0.638, p<0.001).

Figures 7&8 provide a visualization of each of the adjusted test statistics. Figure 7 shows a visualization of adjusted p-values shown in Table 4 while Figure 8 visualizes each of the 12 β terms. For each event, changes in prevalence are represented by the vertical distance between the two regression lines (shown in black) at the event of interest (represented by the dotted line). Changes in incidence are reflected by uneven slopes between the two lines.

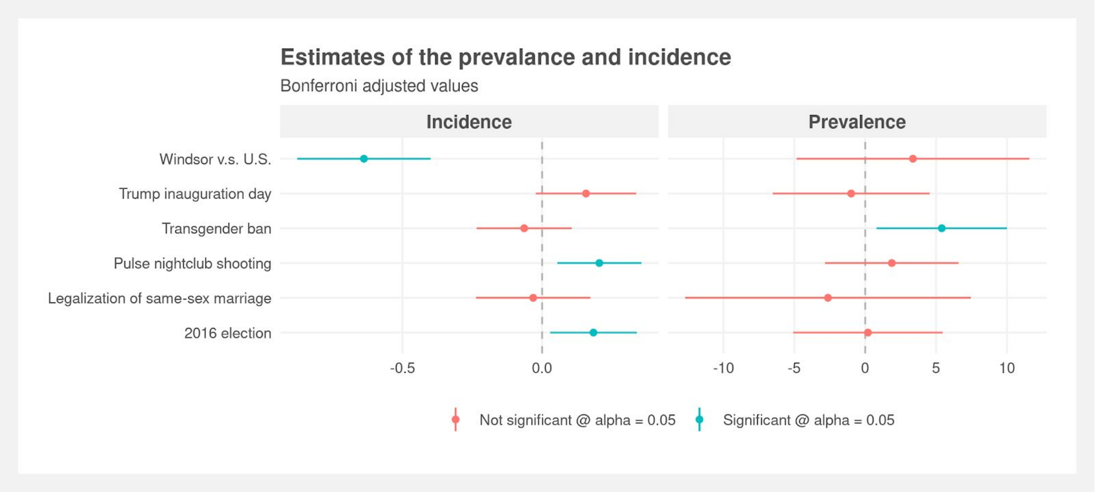

Figure 7: Estimates of incidence and prevalence

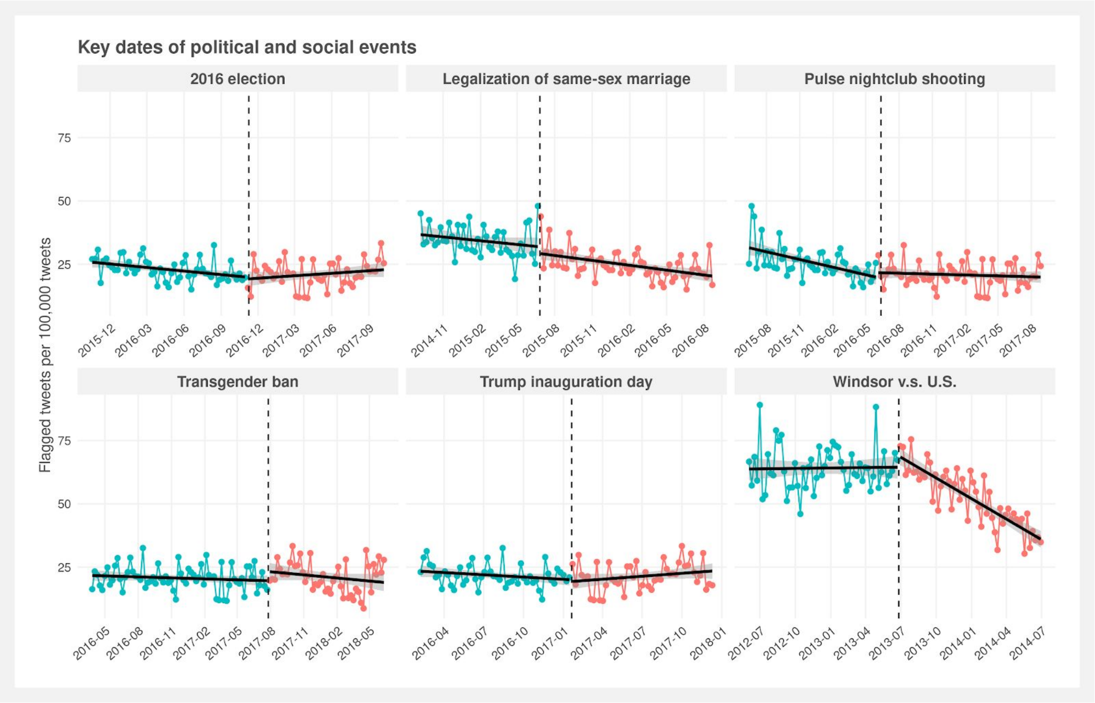

Figure 8: Interrupted time series results

Discussion

Prior work from Siegel and colleagues (2019) suggest that divisive political events in the United States may be associated with slight increases in the frequency of hate speech in Twitter, but quickly dissipate to baseline levels and show no enduring changes in the rate of Twitter-based hate speech. This work expanded on the work of Siegel and colleagues by exploring language explicitly directed at the LGBTQ+ community and by examining multiple politically charged events uniquely relevant to the LGBTQ+ community.

We initially hypothesized that patterns in anti-LGBTQ+ language would follow the results of Siegel and colleagues' (2019) findings on the prevalence and incidence of white-nationalist rhetoric following the 2016 presidential election. Contrary to expectations, the findings reported here contrast the predicted immediate blip followed by a return to baseline. Only a single event, the announcement of the transgender military ban, coincided with an immediate jump in the prevelance of anti-LGBTQ+ language. The remaining five events did not have any visible change immediately around the event.

Changes in immediate prevalence surrounding the event of interest are not the only theoretically important measure. Media cycles extend well beyond the day or week of a specific event and any of the six events may translate to longer, more enduring changes in the presence of anti-LGBTQ+ language on Twitter. We tested the longer term tends following each of the events using interrupted time series. These results provided evidence that the rate of anti-LGBTQ+ language on Twitter increased following Pulse nightclub shooting and the 2016 election but decreased following the decision of US vs Windsor. Prima facie, the finding that some events are associated with increased incidence in anti-LGBTQ+ language while others are associated with reductions in anti-LGBTQ+ language is an inherent contradiction. This contradiction can be explained when examining the broader pattern of the data.

Figure 9 shows the overall proportion of anti-LGBTQ+ tweets with each of the six events of interest plotted as vertical lines. Anti-LGBTQ+ tweets have a distinct peak in mid 2013 followed by a consistent decline that reaches a baseline in mind 216. When looking at the broader context of tweets, the observed changes in incidence observed for the Pulse nightclub shooting (June 2016) and the 2016 presidential election (November 2016) can be explained as the flattening out of a gradual reduction in anti-LGBTQ+ language on Twitter rather than a noticeable increase following the occurrence of political events. Accordingly, the findings suggesting significant changes in the incidence of anti-LGBTQ+ language following the Pulse nightclub shooting and the 2016 election need to be taken in the context of extreme ambiguity. Incorporating prior knowledge of the existing downward trend and leveling out would likely lead to more skeptical findings and it is unlikely that the political events in question fully, or for that matter partially, explain the observed leveling off of anti-LGBTQ+ language.

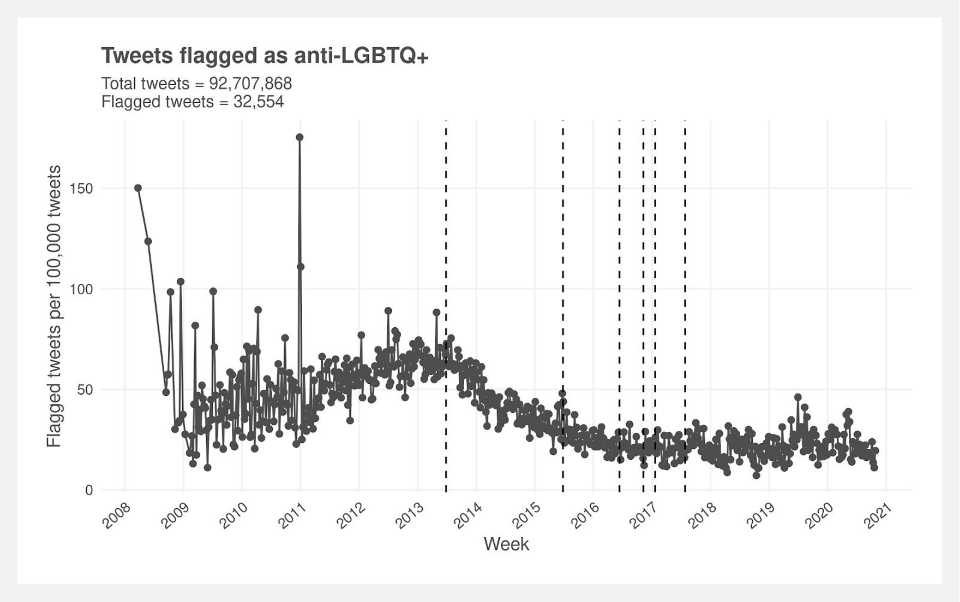

Figure 9: Proportion of tweets identified as anti-LGBTQ+

The most salient shift in LGBTQ+ language occurs midway through 2013, this point is the beginning of a consistent and gradual reduction in anti-LGBTQ+ language that culminates with a leveling off midway through 2016. Notably, this sharp reduction in the incidence of anti-LGBTQ+ language coincided with the Windsor vs U.S. court case. To understand other competing explanations for this sudden shift, we investigate other events that occurred in the same window of time. Investigating potential confounders revealed that in August of 2013, Twitter announced a change in policy allowing users to flag tweets that they found offensive. While the policy was announced in late August of 2013, the implementation was gradual.

Shown in figure 10, the implementation of the new policy aligns with clear reduction in anti-LGBTQ+ language. Space between the decision of Windsor vs U.S. suggests a constant level of anti-LGBTQ+ speech that only begins to change following the introduction of the novel Twitter policy. This realization adds further skepticism to the observed changes in slope for the two events occurring in 2016. Skeptical priors suggest that the incidence in anti-LGBTQ+ tweets was already reducing and reached a new natural baseline around 2016 after the policy has been universally implemented.

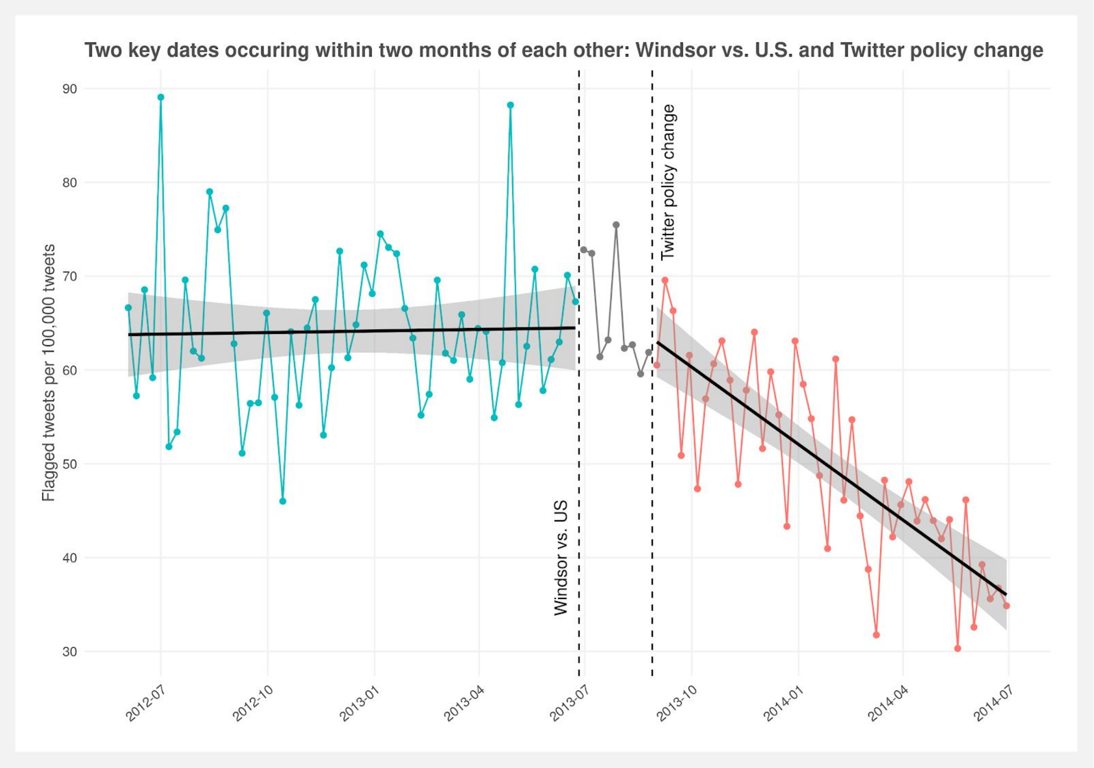

Figure 10: Interrupted time series of pre/post July 2013

Limitations

Aside from the confounding Twitter policy introduced in 2013, several notable measurement and methodological limitations remain. The retrospective nature of the data collection is a major limitation of the study. Rather than identifying active users and tracking anti-LGBTQ+ in a prospective manner, we relied on retrospectively identifying instances of anti-LGBTQ+ language from the past. The sampling strategy we used to obtain past user tweets introduces a potentially biased sample that over-samples users who had either joined the site early on. The limitation of only being able to collect the latest 3,240 tweets from a user biases the sample to more recent tweets. The two combine together to produce the bimodal distribution of tweets by date. While we do not believe this sampling bias is correlated with a propensity to use more anti-LGBQ+ language, we can not formally test this assumption.

After obtaining 92 million distinct tweets from 156,027 users we relied on a content dictionary to screen for potential instances of anti-LGBTQ+ language. This approach follows the framework established by Siegel et al (2019), however, this technique relies heavily on the assumption that the content dictionary is a comprehensive collection of anti-LGBTQ+ language. Terms not included in the content dictionary would not have been picked up by the screener and, subsequently, would not have been included in the analysis. Moreover, language evolves over time and novel forms of hate-speech and derogatory language are fluid. The use of the content dictionary applies a cross-sectional set of terms to a longitudinal dataset. While we believe the content dictionary is broad enough to cover the evolution of anti-LGBTQ+ language over time, this an implicit assumption in the data collection process.

We used the open text field provided by Twitter to limit the sample of tweets to US users. The sample includes users who listed their location as variations of “United States,” states’ names or abbreviations, or included one of the top 1,000 US cities by population. While this approach helped refine the sample to the United States context, users who did not provide a location or were dishonest about their location would have been excluded. Similarly, users who were outside of the United States but set their location to a United States city (or the United States as a country) would have been improperly included. Users are not required to list their locations and those who did not were excluded from the analysis. It is possible that there exists a relationship between transparency about one's location and the propensity to use anti-LGBTQ+ language. If this is the case, our sampling strategy may have systematically under- or over-represented the amount of users posting anti-LGBTQ+ content.

The final sample of tweets included content not in english. The content dictionary flagged non-english tweets with words or phrases that resemble english words included in the content dictionary. The research team was unable to develop NLP tools to translate and classify non-english tweets.

While the classification model was effective at classifying out of sample tweets (0.8 AUC), all three machine learning models were not perfect classifiers. In practice, AUC of 0.9 is a desirable target and none of our three models were able to reach this threshold. The predictive power of all three machine learning models, and particularly the RNN, were limited by the low number of labeled tweets in the training set. The research team was only able to code 6,000 tweets to use as training data. Prior studies suggest a training set of 25,000 labeled cases to reach an out of sample prediction AUC of greater than 0.9 (Siegel et al., 2019). We make the strong assumption that there is no relationship between misclassified tweets and time. If this assumption holds, the observed trends in anti-LGBTQ+ language will not be affected by misclassified tweets. However, if the misclassified tweets are correlated with specific times, it is possible that the observed associations between incidence of anti-LGBTQ+ language and identified events may have been based on the classifier.

Conclusion & future research

We were able to detect noticeable trends in the occurrence of anti-LGBTQ+ tweets. Originally, this research aimed to investigate the relationship between divisive political events relevant to the LGBTQ+ community and the occurrence of increased anti-LGBTQ+ language on a popular social media site. An unexpected finding was the potential impact of a Twitter policy change that coincided with a noticeable reduction in the incidence of offensive tweets targeted at the LGBTQ+ community.

Future research aims to expand this exploratory finding by rigorously testing the policy implications of allowing users to report posts they find offensive in nature. This research will also aim to implement improved methodologies by expanding the content dictionary of hate speech and develop a more comprehensive code book to label training tweets for machine learning classifiers. As described, the content dictionary used as the first step of identifying anti-LGBTQ+ tweets was based only off of hatebase.org. A more expansive content dictionary could be created by including language obtained from known sources of anti-LGBTQ+ content such as the message board 4chan or known anti-LGBTQ+ Reddit communities. When coding screened tweets, explicitly separating clear hate speech from ambiguous cases such as colloquial slurs. Moreover, limiting analysis to users who have had a history of posting at least one instance of anti-LGBTQ+ content may reduce variance in the baseline prevalence of anti-LGBTQ+ language and allow for precise statistical analyses. This approach would examine the effect of Twitter policy and political events on a subset of users with a demonstrated propensity to using anti-LGBTQ+ language.

Future analysis could also allow for more precise estimates. The current study only used a single model to classify tweets identified from the content dictionary screener. While we only used predictions from the RNN in this study, continuing work aims to test the potential of ensemble models that combine the predictions of several different models. Such ensemble methods aim to leverage the different strengths of separate classifiers to produce predictions that are more robust than those of a single model. After classifying tweets, more advanced models can be used to quantify patterns in the incidence and prevalence of tweets that are harmful to the LGBTQ+ community. ARIMA methods and Bayesian time-series models present an alternative method to quantify changes in a more noise tolerant manner compared to the RDD and interrupted time series designs used in the present report.

Appendix

References

Siegel, A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., ... & Tucker, J. A. (2019). Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath.

Lewis, D. C., Flores, A. R., Haider-Markel, D. P., Miller, P. R., Tadlock, B. L., & Taylor, J. K. (2017). Degrees of Acceptance: Variation in Public Attitudes toward Segments of the LGBT Community. *Political Research Quarterly*, *70*(4), 861–875. https://doi.org/10.1177/1065912917717352

Harris Insights & Analytics. (2018). A Survey of American Acceptance and Attitudes Toward LGBTQ Americans. https://www.glaad.org/files/aa/Accelerating%20Acceptance%202018.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *ICWSM*.

Gelron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly.

Calonico, S., Cattaneo, M., Farrell, M., & Titiunik, R. (2020) Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs
https://CRAN.R-project.org/package=rdrobust

Dhinakaran, B. C., Nagamalai, D., & Lee, J. K. (2009, June). Bayesian approach based comment spam defending tool. In *International Conference on Information Security and Assurance* (pp. 578-587). Springer, Berlin, Heidelberg.

King, R. K. (2010, June 1). Announcing Snowflake. blog.Twitter.com.
https://blog.twitter.com/engineering/en_us/a/2010/announcing-snowflake.html

Software

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roesslein, J. (2020). Tweepy: Twitter for Python! *URL: https://Github.Com/Tweepy/Tweepy*

Data structures for statistical computing in python, McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (Publisher link).

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell and Rocio Titiunik(2020). rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs. R package version 0.99.9.https://CRAN.R-project.org/package=rdrobust

Hipp, R. D. (2020). SQLite. Retrieved from https://www.sqlite.org/index.html