Measuring Anti-LGBTQ+ Language on Twitter

Joe Marlo, George Perrett, Bilal Waheed

NYU Statistical Consulting
Fall 2020

Motivation

Do LGBTQ+ people (and other minority populations) feel safe living in the U.S?

The highly polarized and divisive tactics used in the 2016 presidential election was a cause for concern about the perceived impact on the general population.

Anecdotal evidence shows that certain minority populations felt increasingly unwelcome.

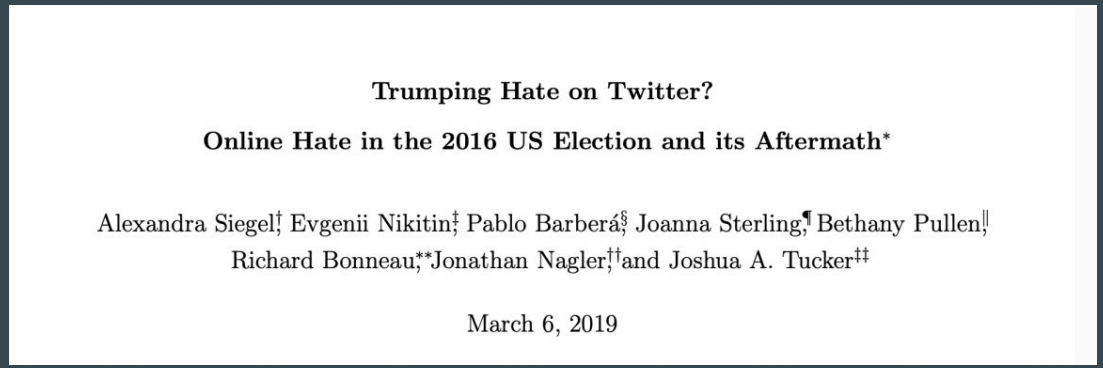
- Is there evidence of this on Twitter at scale?
  - Center of political discord and vector of hate speech
  - Popular platform to express opinions

- Is there a connection between social media, anti-LGBTQ+ language speech, and political events?

Prior Research on Hate Speech

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 the prevalence of anti-LGBTQ+ language.



Trumping Hate on Twitter?

Online Hate in the 2016 US Election and its Aftermath*

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

March 6, 2019

Our 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?

How do we identify anti-LGBTQ+ language?

There isn't an agreed upon definition about what hateful language is, but we know it when we see it.

- Consulted outside resources for guidance on interpretations.
- Previous literature (Davidson, 2017) generally agrees that hateful language is hostile, derogatory language targeted to a specific group of people intended to humiliate, insult or harm members of that group.
- hatebase.org

Our Definition: *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, insult, or harm the person and/or members of the group*

The Data

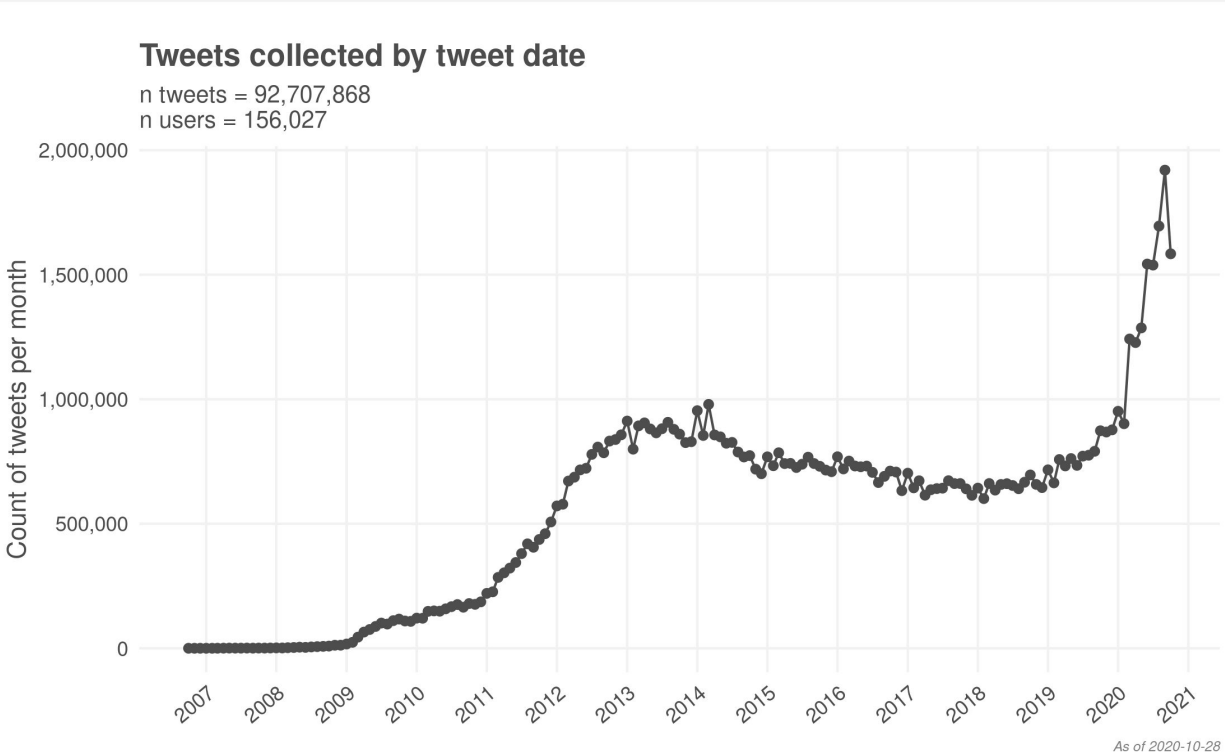
We require **a large number of tweets** to ensure the sample contains enough potential anti-LGBTQ+ language within the periods of interest.

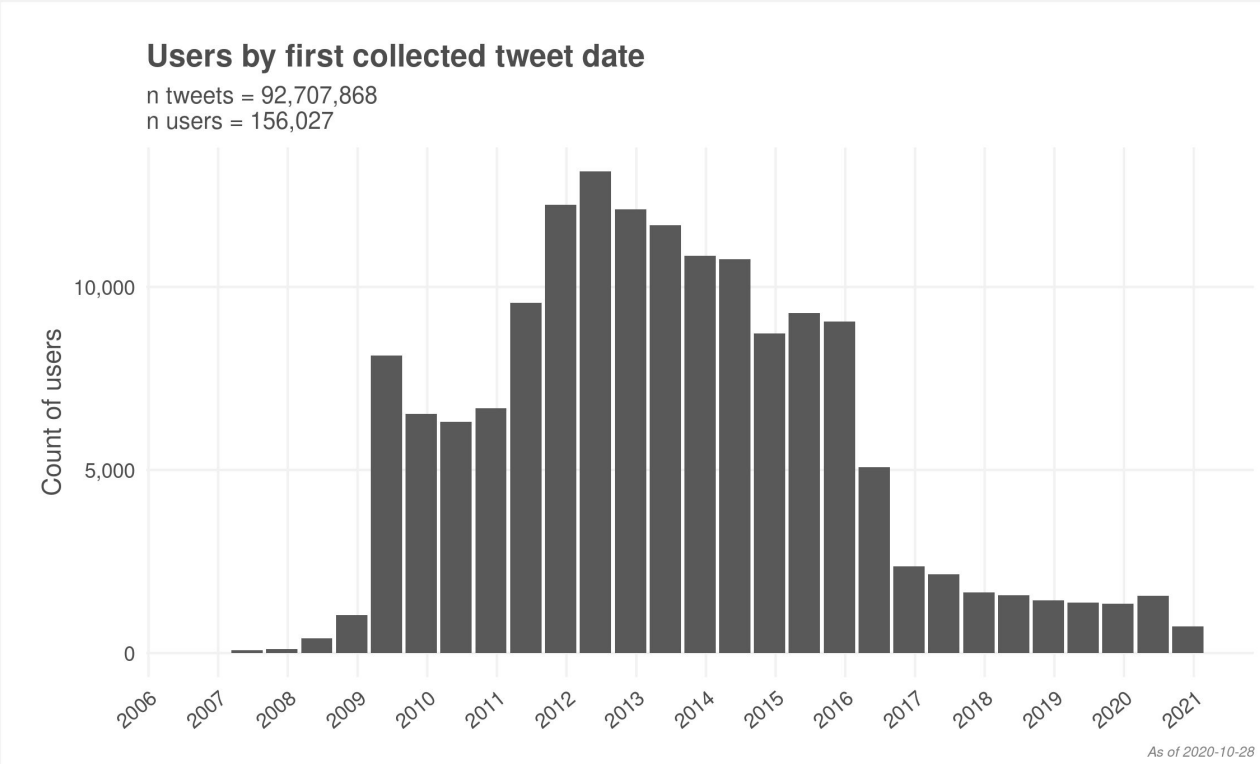
- Early trials indicated that about 0.15% of tweets contain hate words and about 10% of those contain anti-LGBTQ+ language
    - For every identified tweet, we would need approximately 7,000 tweets.

Collected a random sample **over 150,000 U.S. Twitter accounts** and their tweet history, totalling about **92 million tweets.**

Took **25 days of continuous API calls** to build the dataset

- Due to technical restrictions, the sample is biased towards accounts that opened in 2012-2015


Tweets collected by tweet date
n tweets = 92,707,868
n users = 156,027
Count of tweets per month
2,000,000
1,500,000
1,000,000
500,000
0
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
As of 2020-10-28


Users by first collected tweet date
n tweets = 92,707,868
n users = 156,027
Count of users
10,000
5,000
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
As of 2020-10-28

Applying our definition to the dataset

Our definition of hate speech resulted in **identifying over 150,000 tweets** as potential candidates for containing anti-LGBTQ+ language.

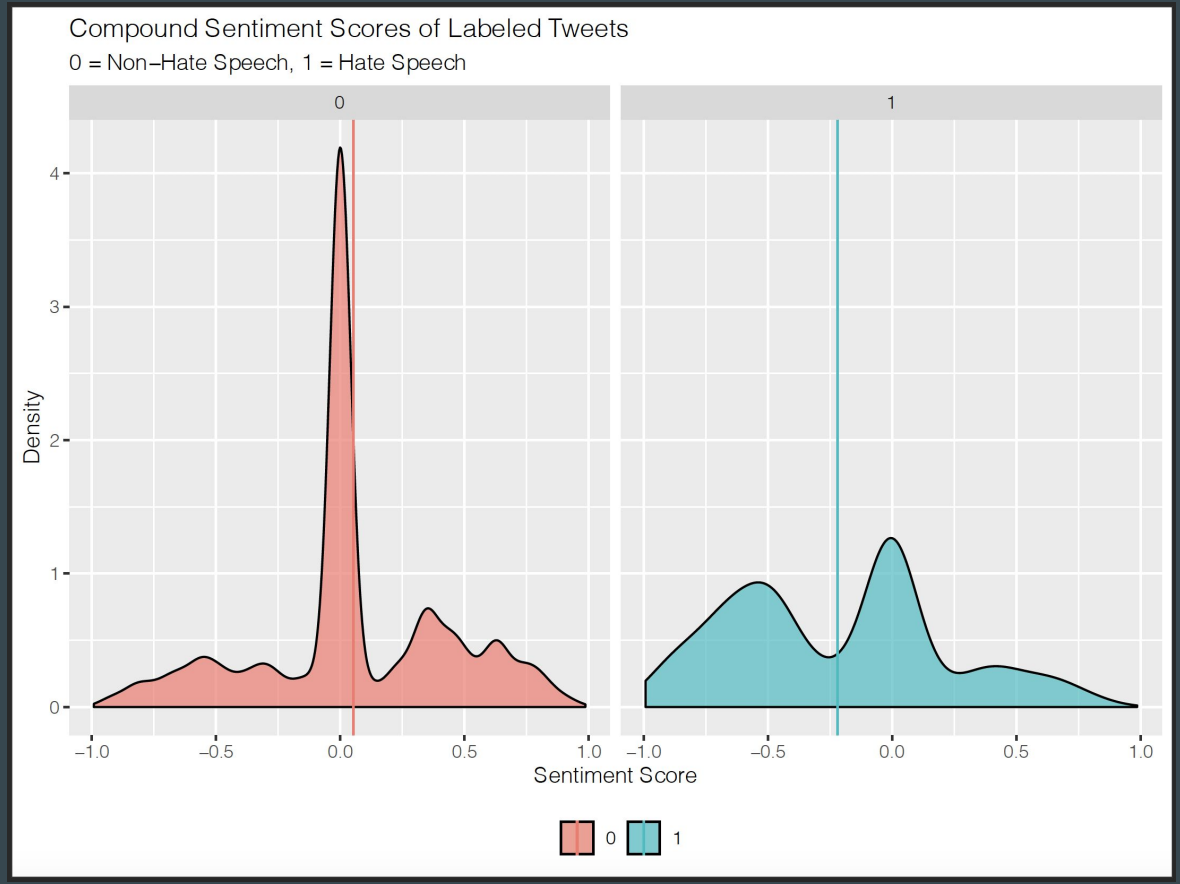
Tweet	Description
“rt @bhand_engineer: @zakirism kuch palo ke liye apne marg se bhatak gaya tha prabhu. aapko vishwaasghat dena humaara maqsat nahi tha. innoc...”	Tweet not in English
“can’t believe i’ve been gay for 23 years and tomorrow is going to be my first time going to pride”	Non-negative LGBTQ+ 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

Delineating between true cases and false flags

We need examples of anti-LGBTQ+ tweets to serve as a training set for our model.

From the 150,000 potential candidates, we randomly **sampled 6,000 tweets and manually labeled** those that explicitly represent anti-LGBTQ+ language

Tweet	Description	Ground truth
“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.	0
“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	0
“when you wanna go out but all your friends are gay af”	More ambiguous case	1
“rt @chefpolohoe: u gay af for lettin dat shit buss all in yo mouth like dat https://t.co/xtlp6s8dri”	Explicit hate speech.	1


Compound Sentiment Scores of Labeled Tweets
0 = Non–Hate Speech, 1 = Hate Speech
0
1
Density
4
3
2
1
0
–1.0
–0.5
0.0
0.5
1.0
Sentiment Score
0
1

Modeling

Logistic Regression

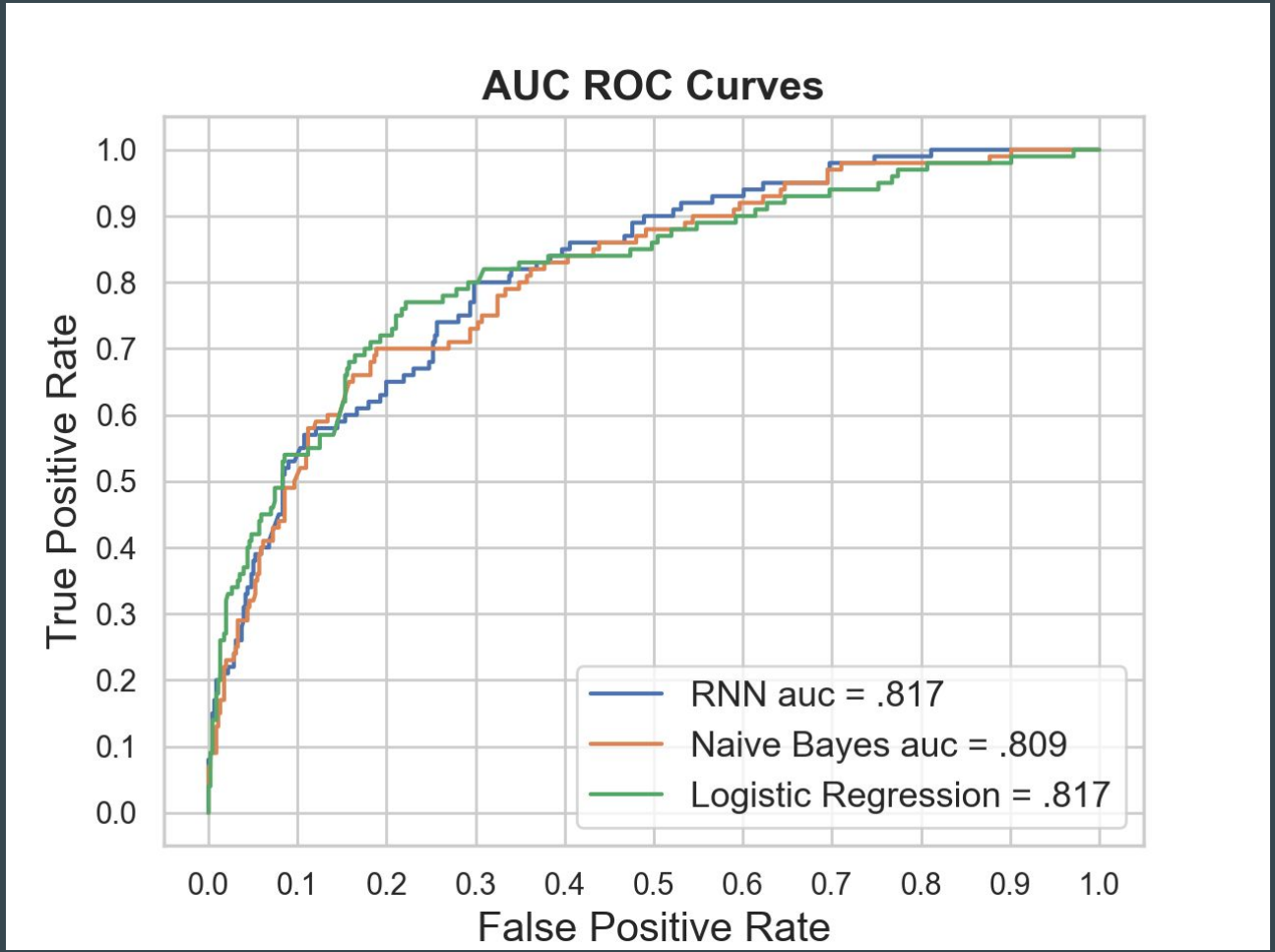
- Allows for word embeddings and estimated probabilities for each word

Naive Bayes (NB)

- Conditional probability estimates from document term matrix

Recurrent Neural Network (RNN)

- Models trained **relationships** between words
- 743,611 trainable parameters
- Incorporates the ability analyze text at the whole tweet level
- Word embeddings

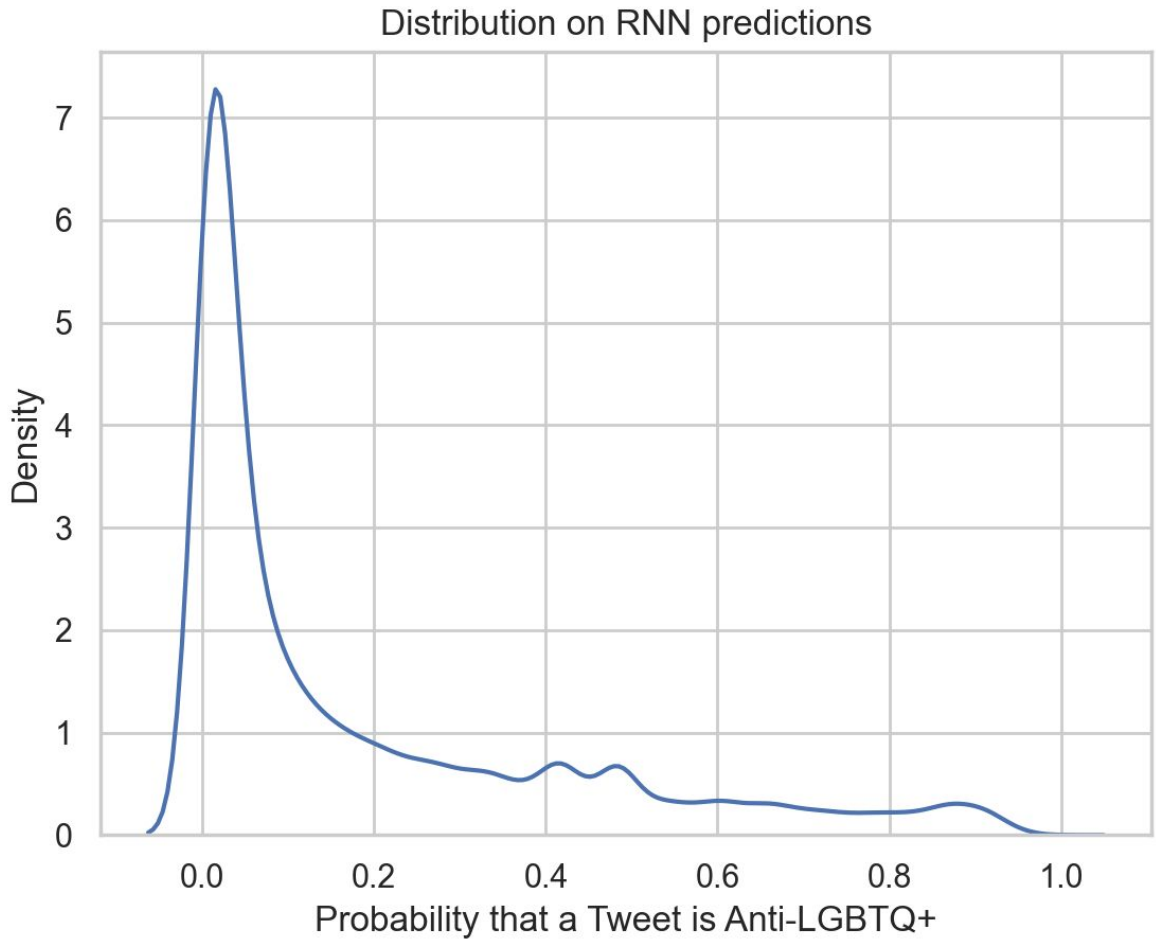

AUC ROC Curves
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
RNN auc = .817
Naive Bayes auc = .809
Logistic Regression = .817

Precision and Recall

Model	Precision	Recall
RNN	.53	.56
Naive Bayes	.66	.39
Logistic Regression	.66	.23

How did the model do in those cases?

Tweet	Description	Ground truth	Model Prob.
“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.	0	0.007
“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	0	0.011
“when you wanna go out but all your friends are gay af”	More ambiguous case	1	.4642
“rt @chefpolohoe: u gay af for lettin dat shit buss all in yo mouth like dat https://t.co/xtlp6s8dri”	Explicit anti LGBTQ+.	1	0.902


Distribution on RNN predictions
Density
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0
Probability that a Tweet is Anti-LGBTQ+

Incidence of anti-LGBTQ+ language

Applied our final model to our entire set (150,000+) of potential candidates of hate speech.

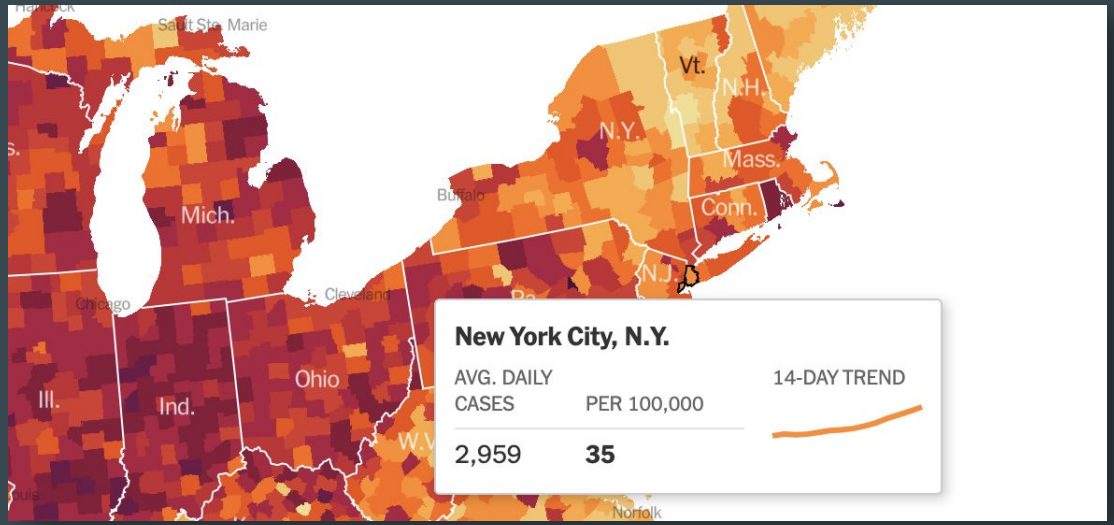
0.035%

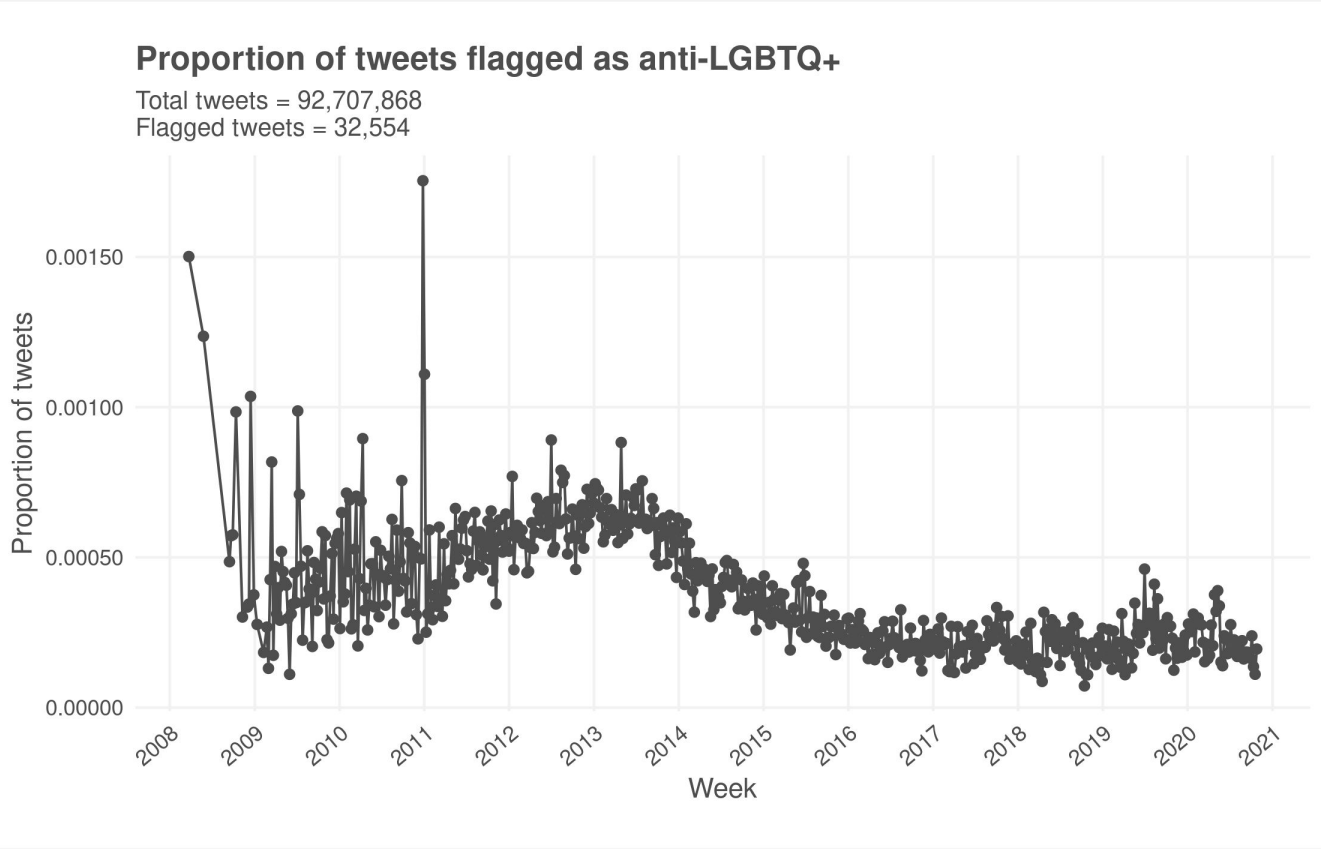
Model classified as anti-LGBTQ+ language

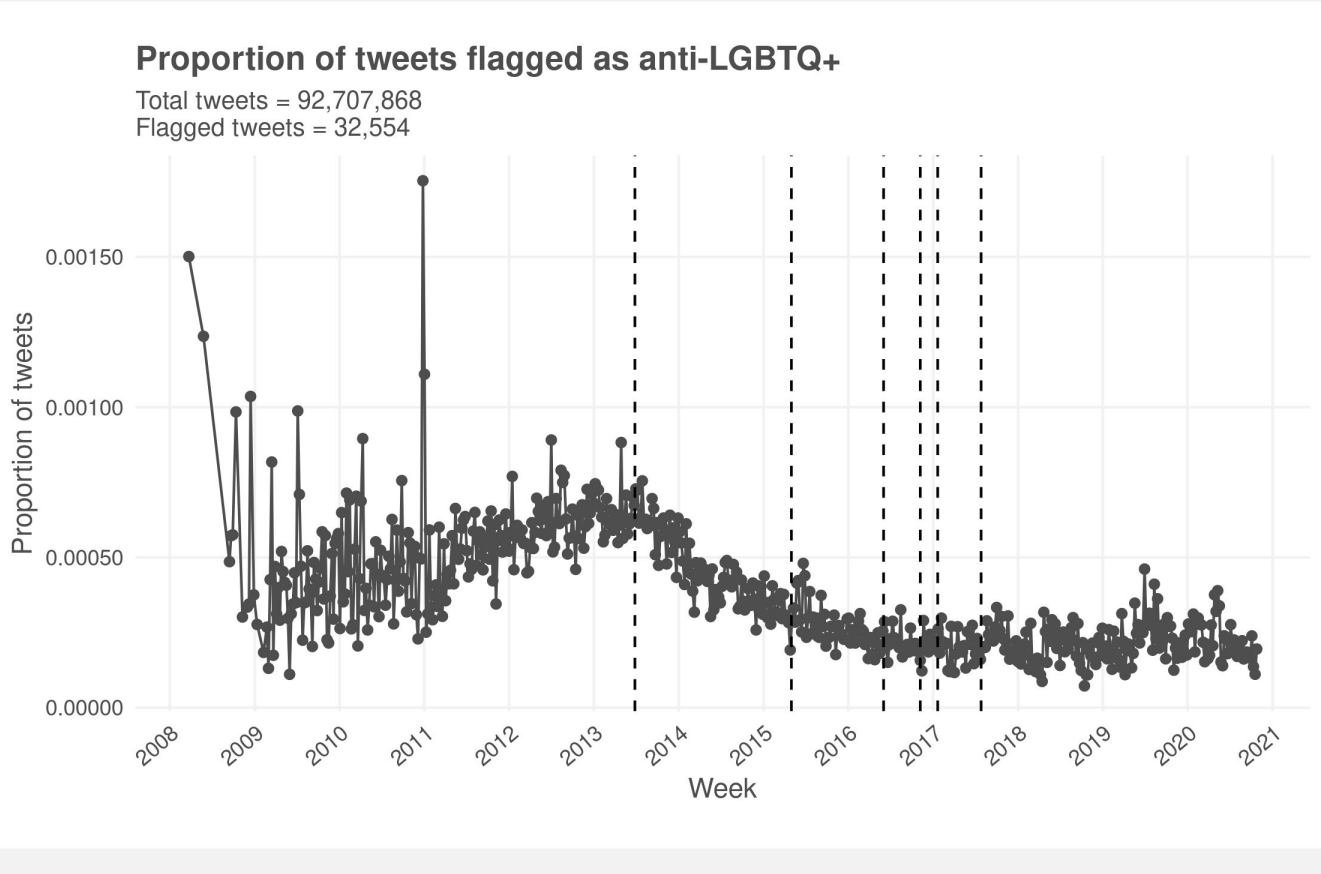
= 35 in 100k tweets

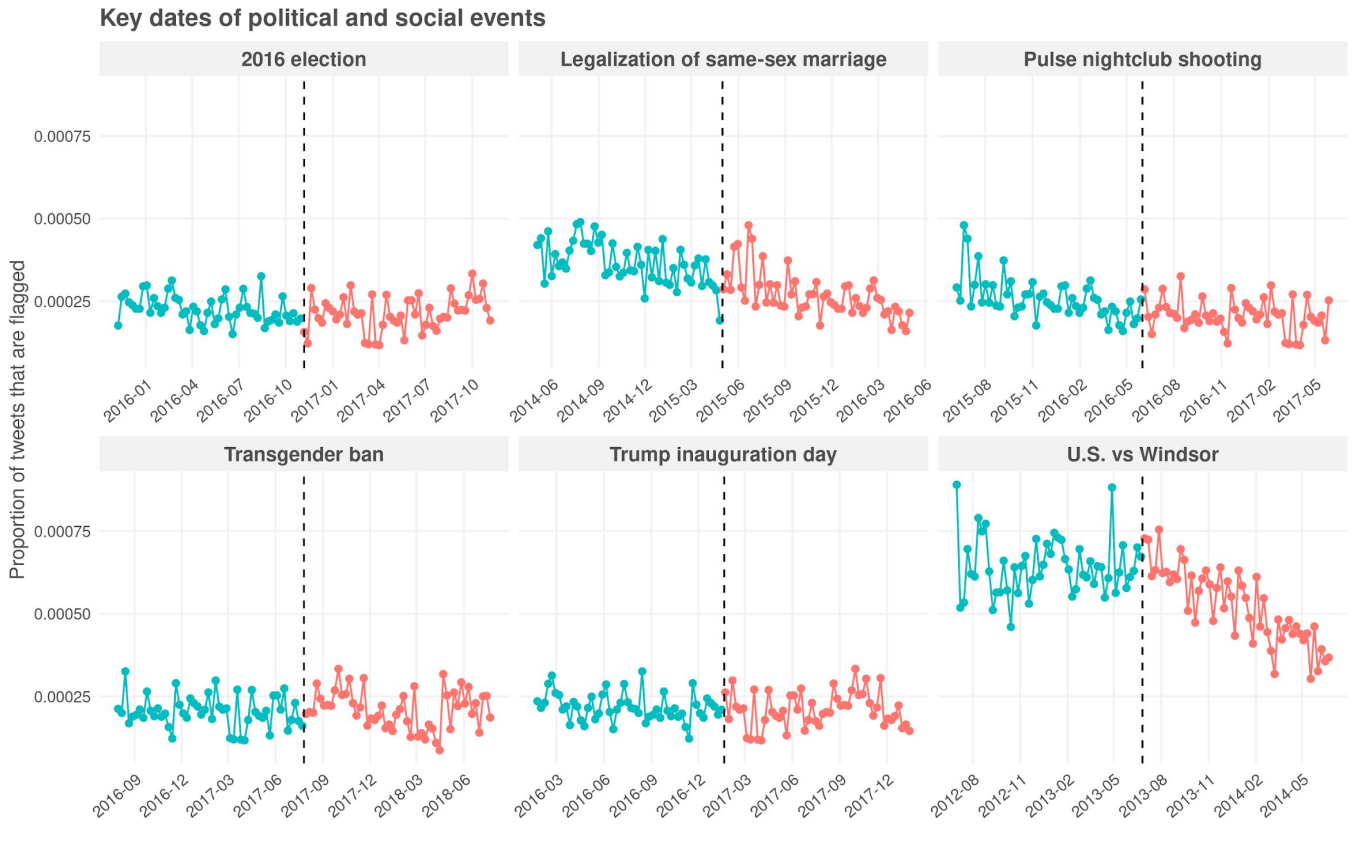
Incidence of anti-LGBTQ+ language

Applied our final model to our entire set (150,000+) of potential candidates of hate speech.

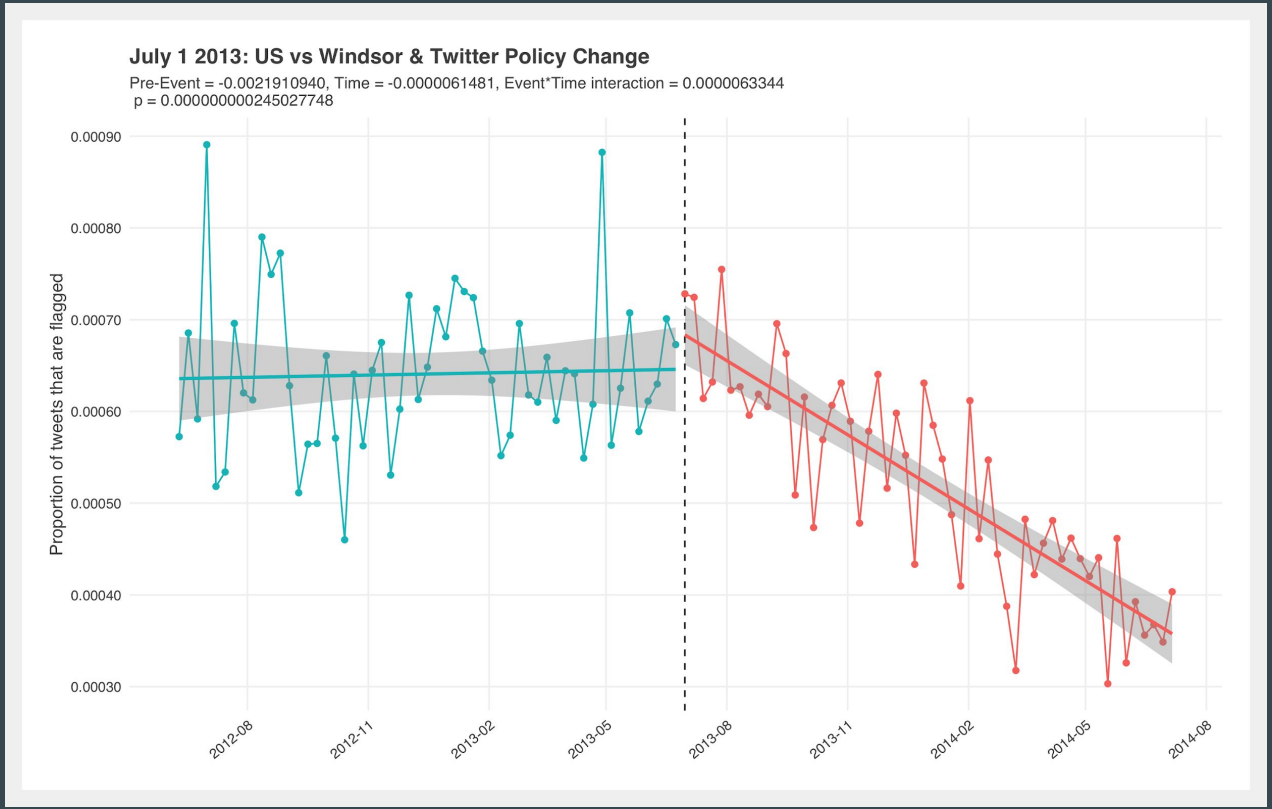



Proportion of tweets flagged as anti-LGBTQ+
Total tweets = 92,707,868
Flagged tweets = 32,554
Proportion of tweets
0.00150
0.00100
0.00050
0.00000
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Week


Proportion of tweets flagged as anti-LGBTQ+
Total tweets = 92,707,868
Flagged tweets = 32,554
Proportion of tweets
0.00150
0.00100
0.00050
0.00000
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Week


Key dates of political and social events
2016 election
Legalization of same-sex marriage
Pulse nightclub shooting
Transgender ban
Trump inauguration day
U.S. vs Windsor
Proportion of tweets that are flagged
0.00075
0.00050
0.00025
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04
2017-07
2017-10
2014-06
2014-09
2014-12
2015-03
2015-06
2015-09
2015-12
2016-03
2016-06
2015-08
2015-11
2016-02
2016-05
2016-08
2016-11
2017-02
2017-05
2016-09
2016-12
2017-03
2017-06
2017-09
2017-12
2018-03
2018-06
2016-03
2016-06
2016-09
2016-12
2017-03
2017-06
2017-09
2017-12
2012-08
2012-11
2013-02
2013-05
2013-08
2013-11
2014-02
2014-05

Interrupted time series analysis



Final thoughts

Limitations and challenges

- **Grant writing** is an essential part of research in academia
- **Research planning** and **experimental design** matters
- Data gathering is time consuming and **sometimes compromises need to be made** in the sampling plan
    - Use Twitter streaming API prospectively
- **Breadth of technical skills** is important to processing a large amount of data
    - Importance of SQL, Scala, Python, Bash, etc.

Conclusion

- There is a measurable level of anti-LGBTQ+ language on Twitter that could be influencing peoples' lives.
- Early exploratory finding suggests companies may be able to influence the prevalence of hate-speech and negative content that targets marginalized populations.

Thank you!
Questions? Comments?

Find our work here: github.com/joemarlo/hate-speech