Trumping Hate on Twitter?
Online Hate Speech and White Nationalist Rhetoric in the 2016 US Election Campaign and its Aftermath

Alexandra Siegel
PhD Candidate, New York University
Graduate Research Associate, NYU SMaPP lab

Co-authors: Evgenii Nikitin, Pablo Barberá Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, Joshua Tucker

Text as Data Conference, Princeton, 2017

Hate Speech – Motivation

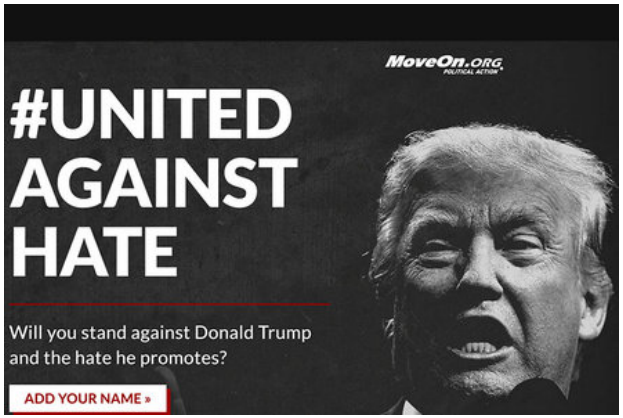
POLITICS SPECIAL REPORTS | Mon Nov 7, 2016 | 10:46pm EST

Hate speech seeps into U.S. mainstream amid bitter campaign

NEWS DESK

HATE ON THE RISE AFTER TRUMP'S ELECTION

By Alexis Okeowo November 17, 2016



DEMOCRACY & GOVERNMENT

Donald Trump and the Escalation of Hate

A number of civil-rights organizations have spoken out about the rise of hate speech and violent threats by groups and individuals who support the presumptive Republican presidential nominee.

BY KARIN KAMP | JUNE 15, 2016

'Massive rise' in hate speech on Twitter during presidential election

Jessica Guynn , USA TODAY Published 5:00 p.m. ET Oct. 21, 2016 | Updated 7:00 p.m. ET Oct. 23, 2016

Research Questions and Basic Definitions

To what extent did online hate speech and white nationalist rhetoric on Twitter increase over the course of Donald Trump's 2016 campaign and in the aftermath of his election?

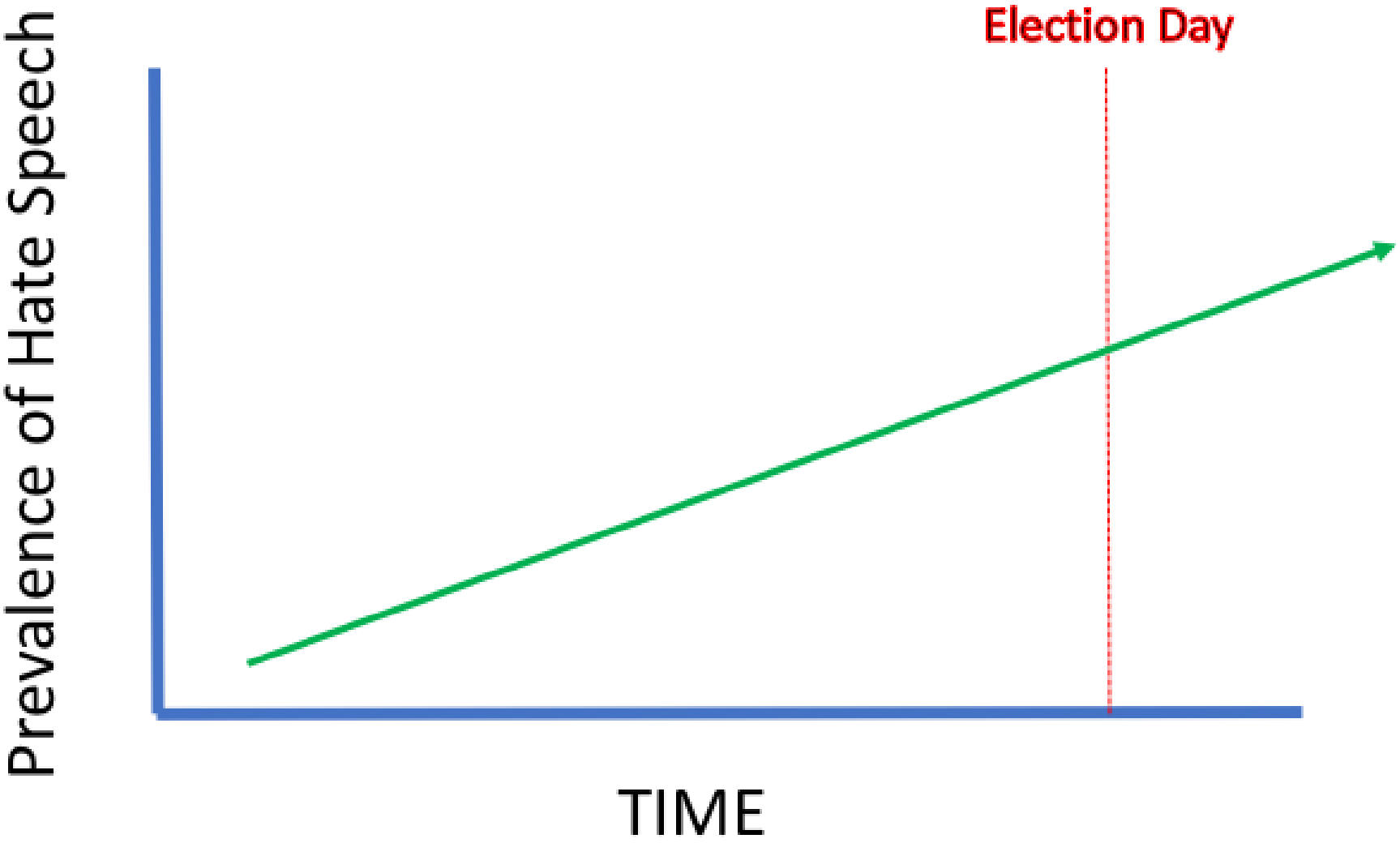
Research Questions and Basic Definitions

To what extent did online hate speech and white nationalist rhetoric on Twitter increase over the course of Donald Trump's 2016 campaign and in the aftermath of his election?

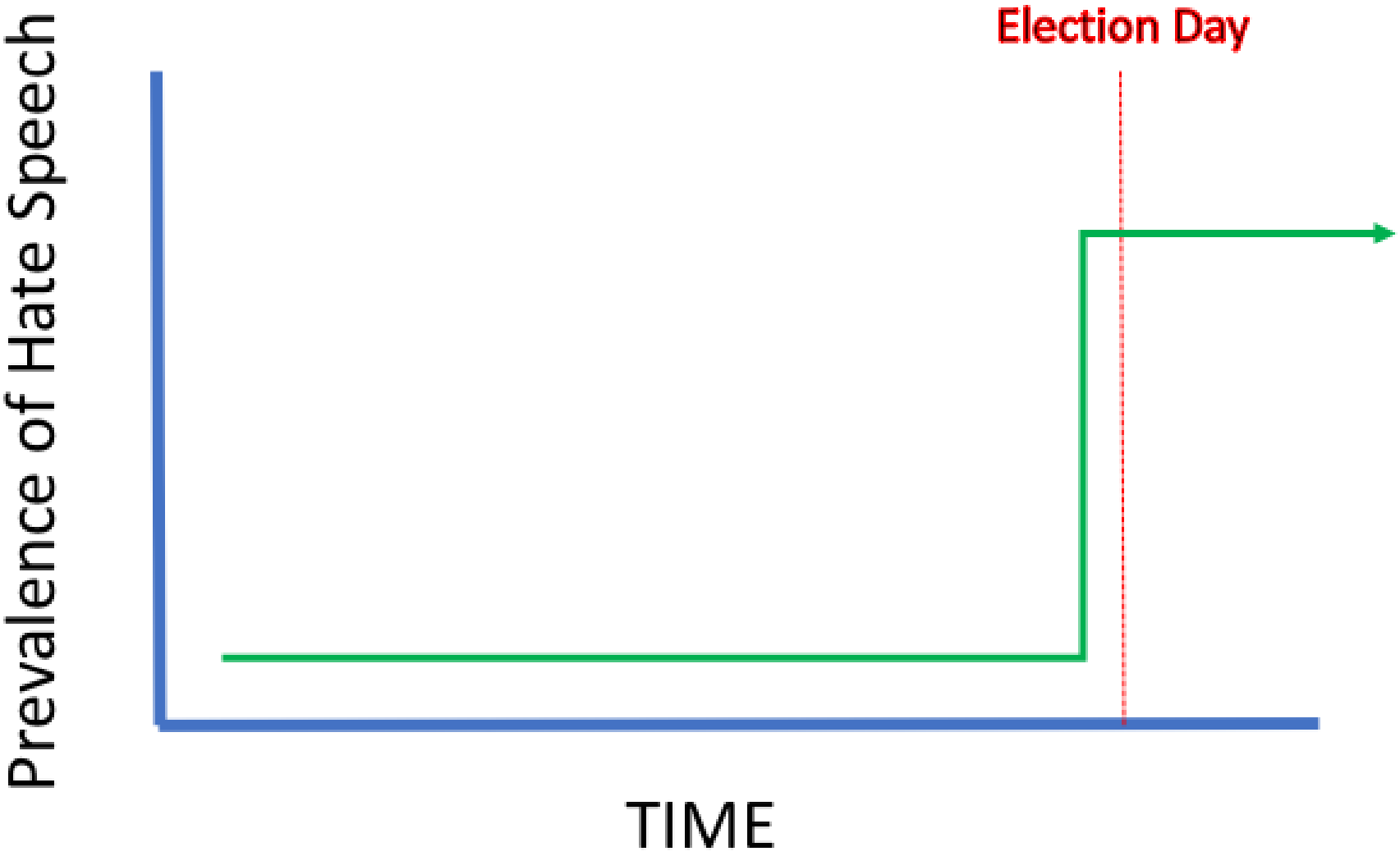
Hate Speech: Bias-motivated, hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics, especially when the group or individual are unnecessarily labeled (Cohen-Almagor 2011).

White Nationalist Language: Rhetoric or content that praises known white-nationalist groups, shows excessive pride in the white race, espouses white supremacist or white separatist ideologies, or focuses on the alleged inferiority of nonwhites (Fording 2014).

Hypothesis 1: Increasing Over Campaign



Hypothesis 2: Increase After Election



Tweet Examples (WARNING: OFFENSIVE LANGUAGE)

Tweet Examples (WARNING: OFFENSIVE LANGUAGE)

Hatespeech Tweet Examples:

- @WhitePriinces: Cant wait for donald trump to send all the monkey looking niggers back home to mexico
- RT @Bidenshairplugz: Fucking Jews, calling Donald Trump "violent" while giving negroes and Muslims a pass on their actual violence.
- Mexiscum parasites want to destroy America in every way they can. Fucking vermin. # BuildTheWall # DeportThemAll https://t.co/bDwM4r2Jug

Tweet Examples (WARNING: OFFENSIVE LANGUAGE)

Hatespeech Tweet Examples:

- @WhitePriinces: Cant wait for donald trump to send all the monkey looking niggers back home to mexico
- RT @Bidenshairplugz: Fucking Jews, calling Donald Trump "violent" while giving negroes and Muslims a pass on their actual violence.
- Mexiscum parasites want to destroy America in every way they can. Fucking vermin. # BuildTheWall # DeportThemAll https://t.co/bDwM4r2Jug

White Nationalist Tweet Examples:

- RT @WhiteGenociders: # DonaldTrump Peaceful White nationalists protect beauty, family, and land, # AntiWhites want to destroy those things
- RT @WhiteGenocideWN: Donald Trump means we don't ever have to apologize for being white ever again # NPI # AltRight @LadyAodh
- @HillaryClinton What is CHASING DOWN every last White person, assimilating them with nonwhites and calling it 'Diversity?...# whitegenocide

Data

Political Twitter:

- Dataset of over 150 million tweets referencing Hillary Clinton collected between June 17, 2015 and June 15, 2017.
- Dataset of over 600 million tweets referencing Donald Trump collected between June 17, 2015 and June 15, 2017.
    - @realDonaldTrump, "trump", @HillaryClinton, "hillary", "clinton", #ImWithHer, #MAGA etc.

Data

Political Twitter:

- Dataset of over 150 million tweets referencing Hillary Clinton collected between June 17, 2015 and June 15, 2017.
- Dataset of over 600 million tweets referencing Donald Trump collected between June 17, 2015 and June 15, 2017.
  - @realDonaldTrump, "trump", @HillaryClinton, "hillary", "clinton", #ImWithHer, #MAGA etc.

Random Sample of American Twitter Users:

- Dataset of about 400 million tweets sent by a random sample of 500,000 American Twitter users collected between June 17, 2015 and June 15, 2017.
  - Numeric ID with random number generator
  - Check for US location

Measuring Online Hate Speech and White Nationalist Language

- Primary Analysis : Dictionary based method + Naive Bayes classifiers to remove false positives

Measuring Online Hate Speech and White Nationalist Language

- Primary Analysis : Dictionary based method + Naive Bayes classifiers to remove false positives
- Robustness Test: Method measuring semantic similarity of tweets and explicitly discriminatory or white nationalist **subreddits**

Measuring Online Hate Speech and White Nationalist Language

- Primary Analysis : Dictionary based method + Naive Bayes classifiers to remove false positives
- Robustness Test: Method measuring semantic similarity of tweets and explicitly discriminatory or white nationalist **subreddits**
- Key point: both methods allow us to measure *change over time* in prevalence of hate speech or white nationalist language

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)
3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)
3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)
2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)
3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:
  - Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:
  - Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump
  - RT @snowangja: Donald Trump is the type to call every east asian people as ching chong too. I'm not shocked https://t.co/zFHwPLmsr9

Dictionary-based Hate Speech Detection on Twitter

1. Create dictionaries of slurs and terms from existing dictionaries of hate speech and white nationalist rhetoric (Hatebase, Racial Slur Database, ADL) $\rightarrow$ (4,477 terms, including variations)

2. Remove terms that are primarily not used as hate speech in a random sample of our Political Twitter dataset. $\rightarrow$ (e.g. pizza, newspaper, soak, taco)$\rightarrow$ (538 terms)

3. Add common Twitter specific terms observed in a random sample of our Political Twitter dataset, and Reddit-specific terms $\rightarrow$ (247 + 268 terms)

Problems with Dictionary Methods:

- Term can be part of a Twitter handle: @angrybitch
- Dictionary terms can be parts of other words: spicy
- Dictionary terms can be homonyms: "a chink in his armor"
- Examples of Anti-Hate Speech that include dictionary terms:
  - Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump
  - RT @snowangja: Donald Trump is the type to call every east asian people as ching chong too. I'm not shocked https://t.co/zFHwPLmsr9
  - RT @ShaunKing: This just happened in Indiana. "Fuck you nigger bitch. Trump is going to deport you back to Africa." Day 1 of Donald

Supervised Classification:

- Trained undergraduates and crowd-source coders on Crowdflower coded a random sample of 25,000 tweets (each tweet coded by 3 people) containing hate speech OR white nationalist rhetoric terms identified using our dictionary method.

Supervised Classification:

- Trained undergraduates and crowd-source coders on Crowdflower coded a random sample of 25,000 tweets (each tweet coded by 3 people) containing hate speech OR white nationalist rhetoric terms identified using our dictionary method.
  - Does this tweet contain hate speech? (yes or no)
  - Does this tweet contain white nationalist rhetoric? (yes or no)
  - Instructions contained detailed definitions and examples.
  - Test questions were used to weed out ineffective coders.

Supervised Classification:

- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.

Supervised Classification:

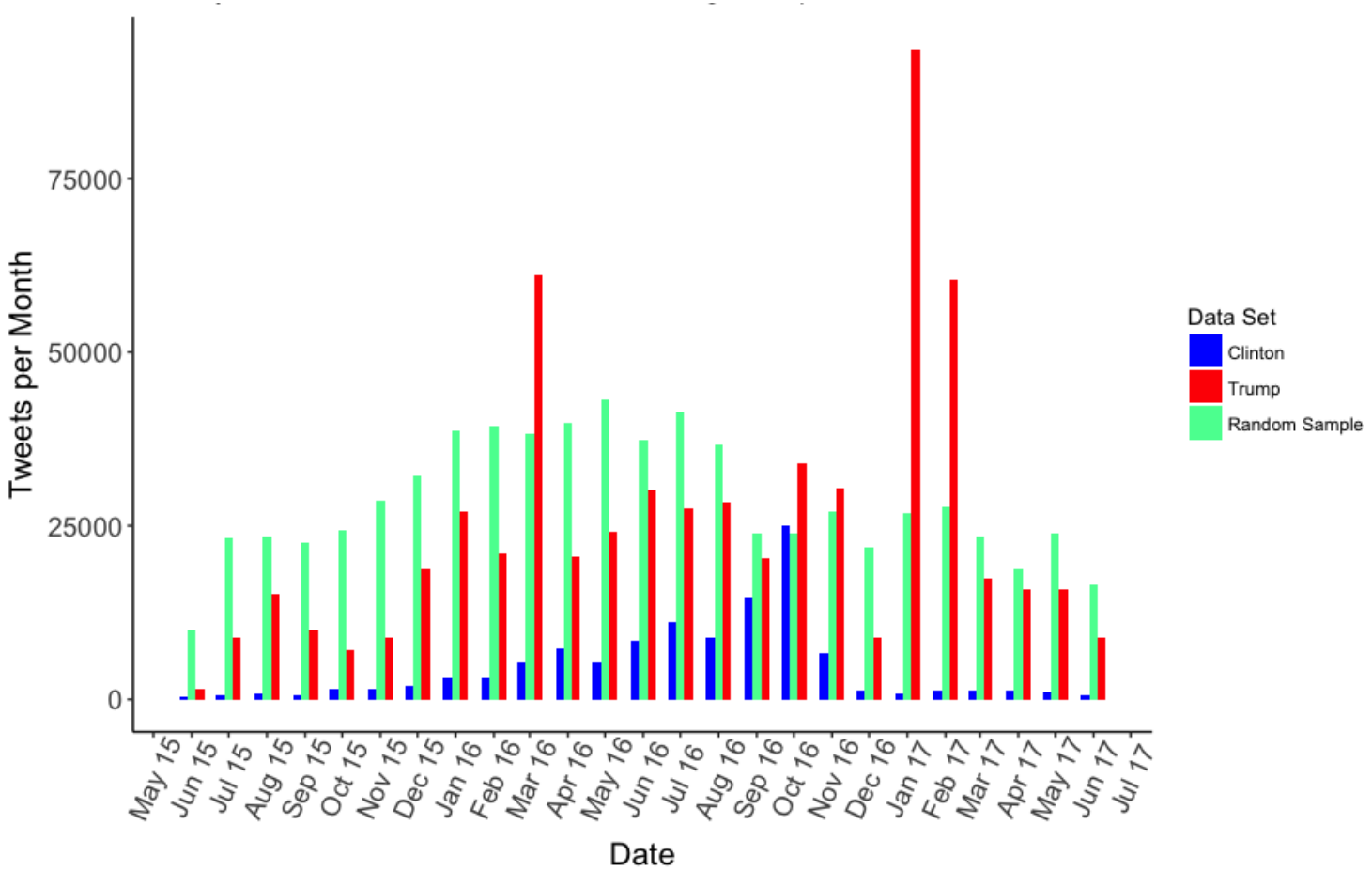
- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.
- Trained two Naive Bayes classifiers (a hate speech and a white nationalist rhetoric classifier).

Supervised Classification:

- According to human coders, fewer than half of the tweets identified *by the dictionary method* in our random sample contained hate speech or white nationalist language.
- Trained two Naive Bayes classifiers (a hate speech and a white nationalist rhetoric classifier).
- We measure the popularity of hate speech and white nationalist rhetoric (WNR) as:
  - The daily proportion of tweets containing hate speech or WNR in each of our datasets.
  - The daily proportion of unique users tweeting hate speech or WNR in each of our datasets.

A look at the data...

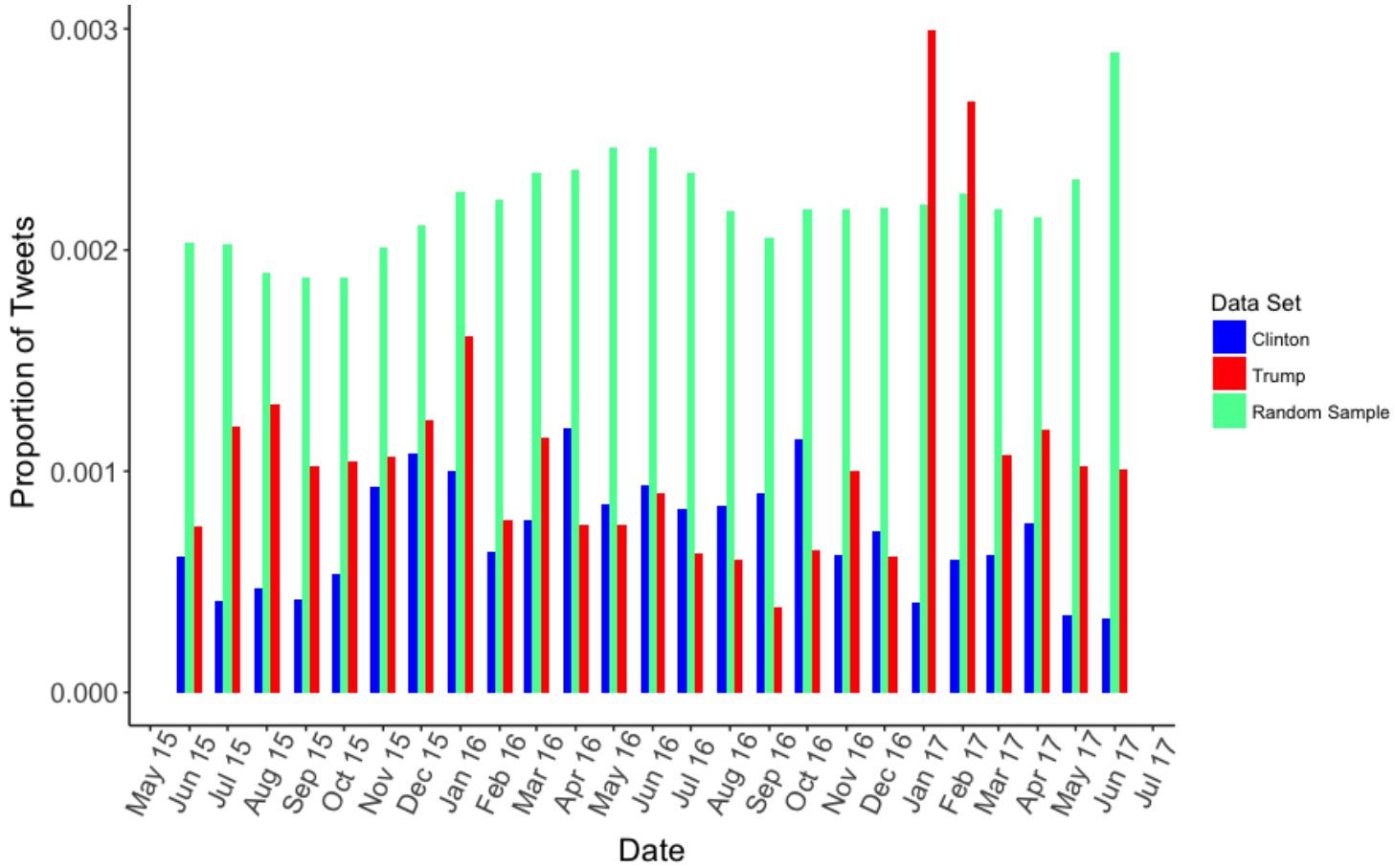
Figure : Monthly Volume of Hate Speech Tweets (All Datasets)



This figure shows the monthly volume of hate-speech tweets in the Clinton, Trump, and random sample datasets of tweets.

A look at the data...

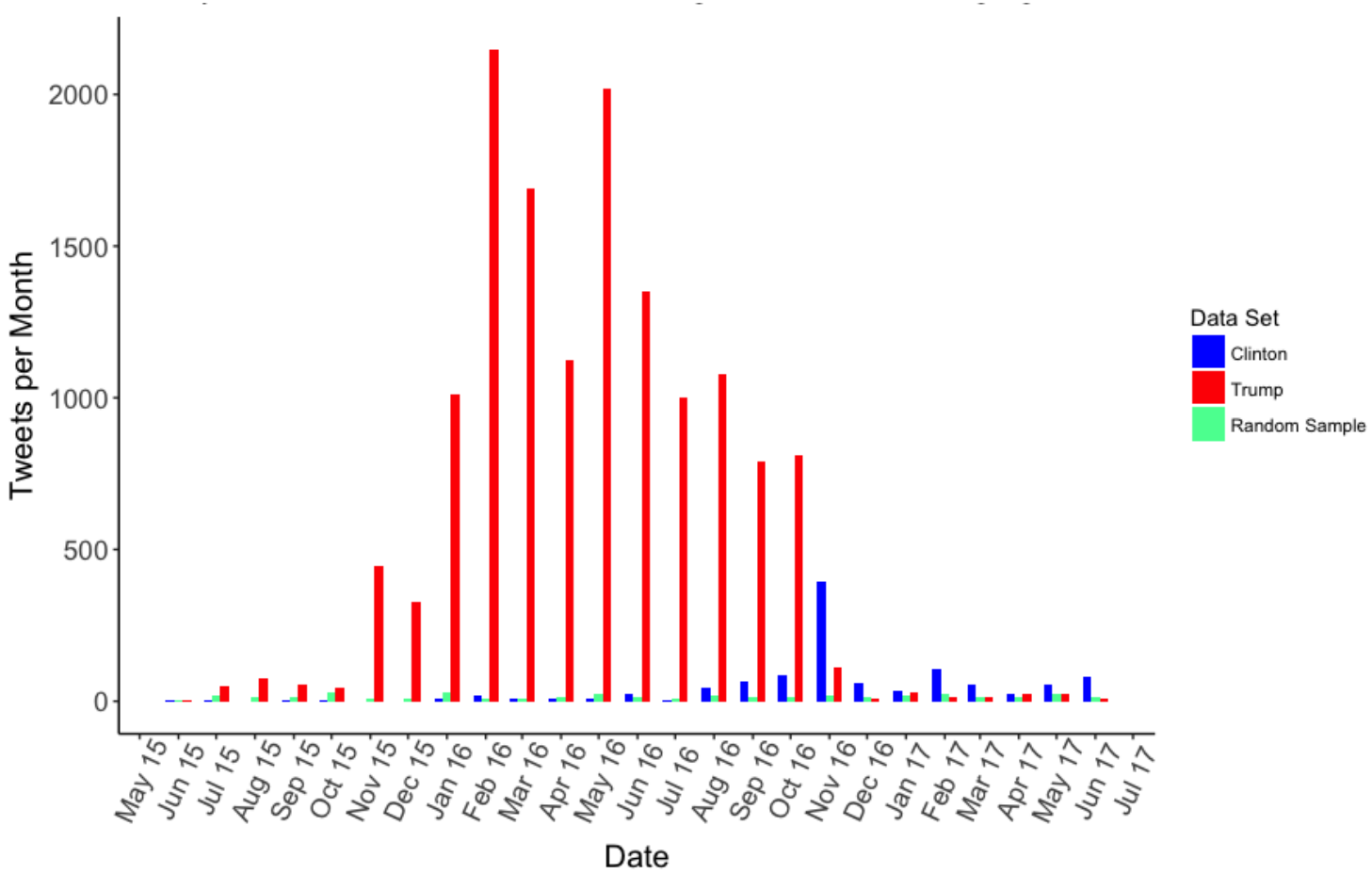
Figure : Monthly Proportion of Hate Speech Tweets (All Datasets)



This figure shows the monthly proportion of hate-speech tweets in the Clinton, Trump, and random sample datasets.

A look at the data...

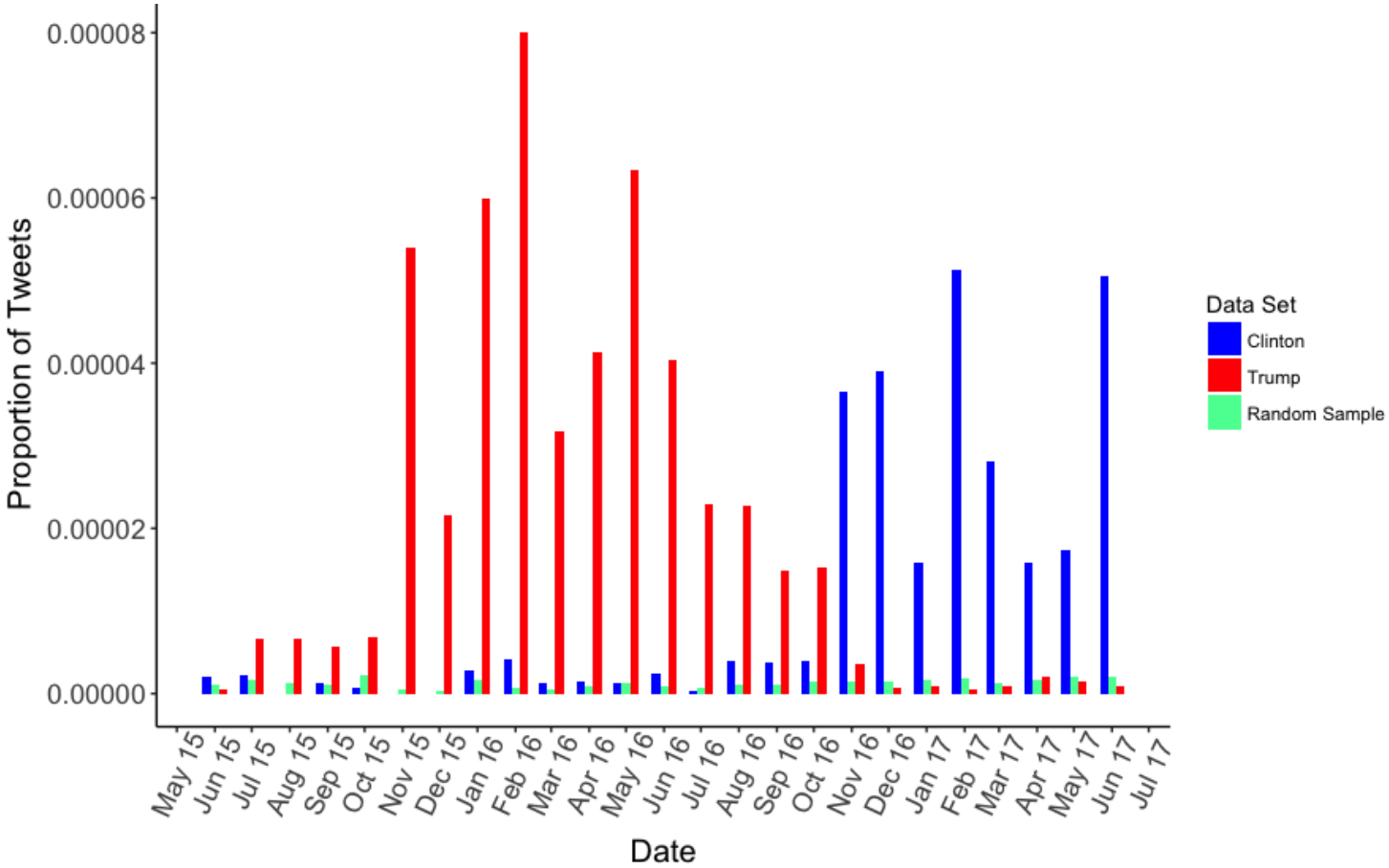
Figure : Monthly Volume of White Nationalist Tweets (All Datasets)



This figure shows the monthly volume of WNR tweets in the Clinton, Trump, and random sample datasets of tweets.

A look at the data...

Figure : Monthly Proportion of White Nationalist Tweets (All Datasets)



This figure shows the monthly proportion of WNR tweets in the Clinton, Trump, and random sample datasets of tweets.

A look at the data...

Hate Speech Types: Anti-Asian, Anti-Black, Anti-Immigrant, Anti-Latino, Anti-Muslim, Anti-Semitic, Homophobic, Misogynistic

Figure : **Daily Proportion of Hate Speech by Topic in Trump Dataset**

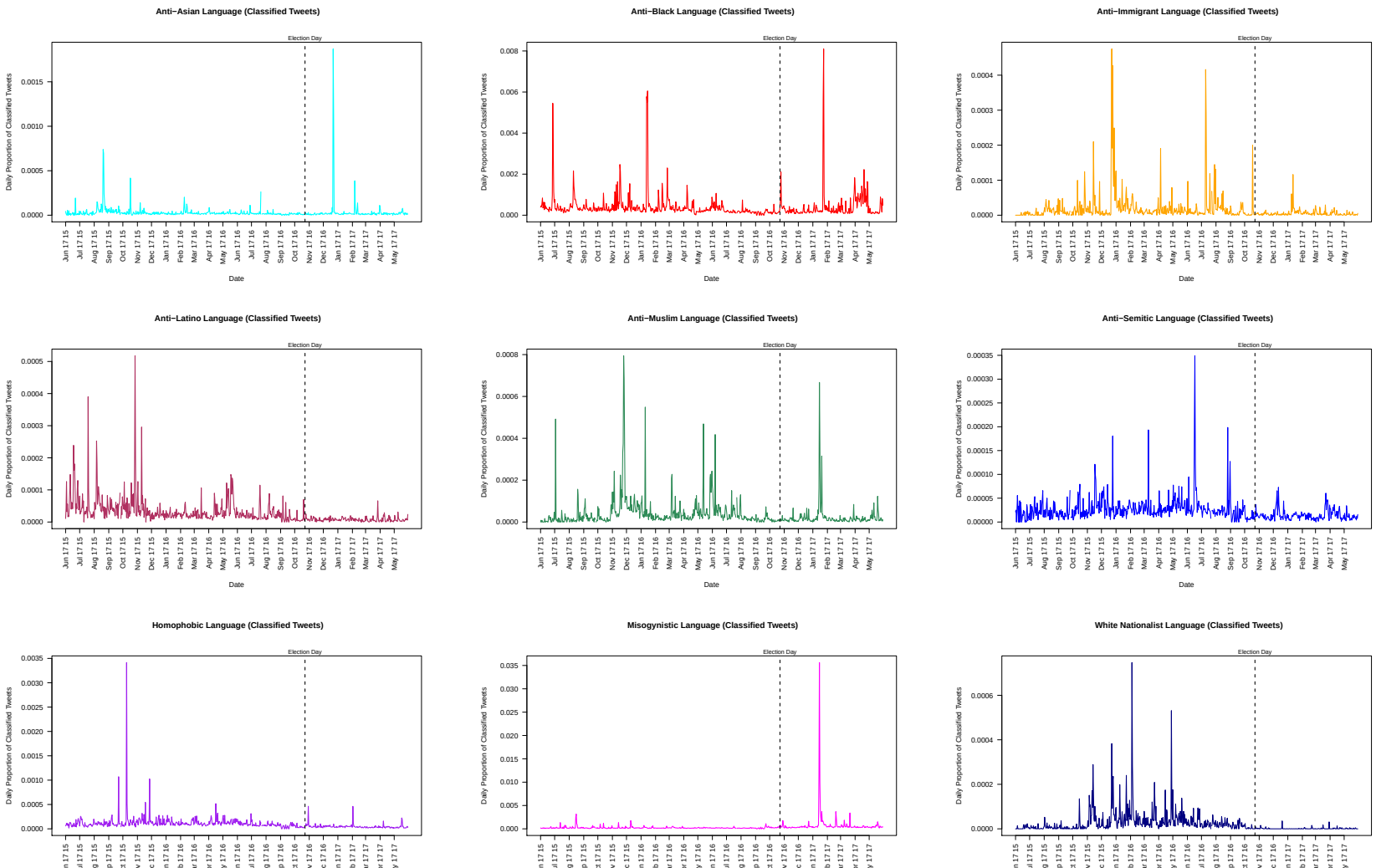


Figure : **Daily Proportion of Anti-Muslim Tweets in Trump Dataset**

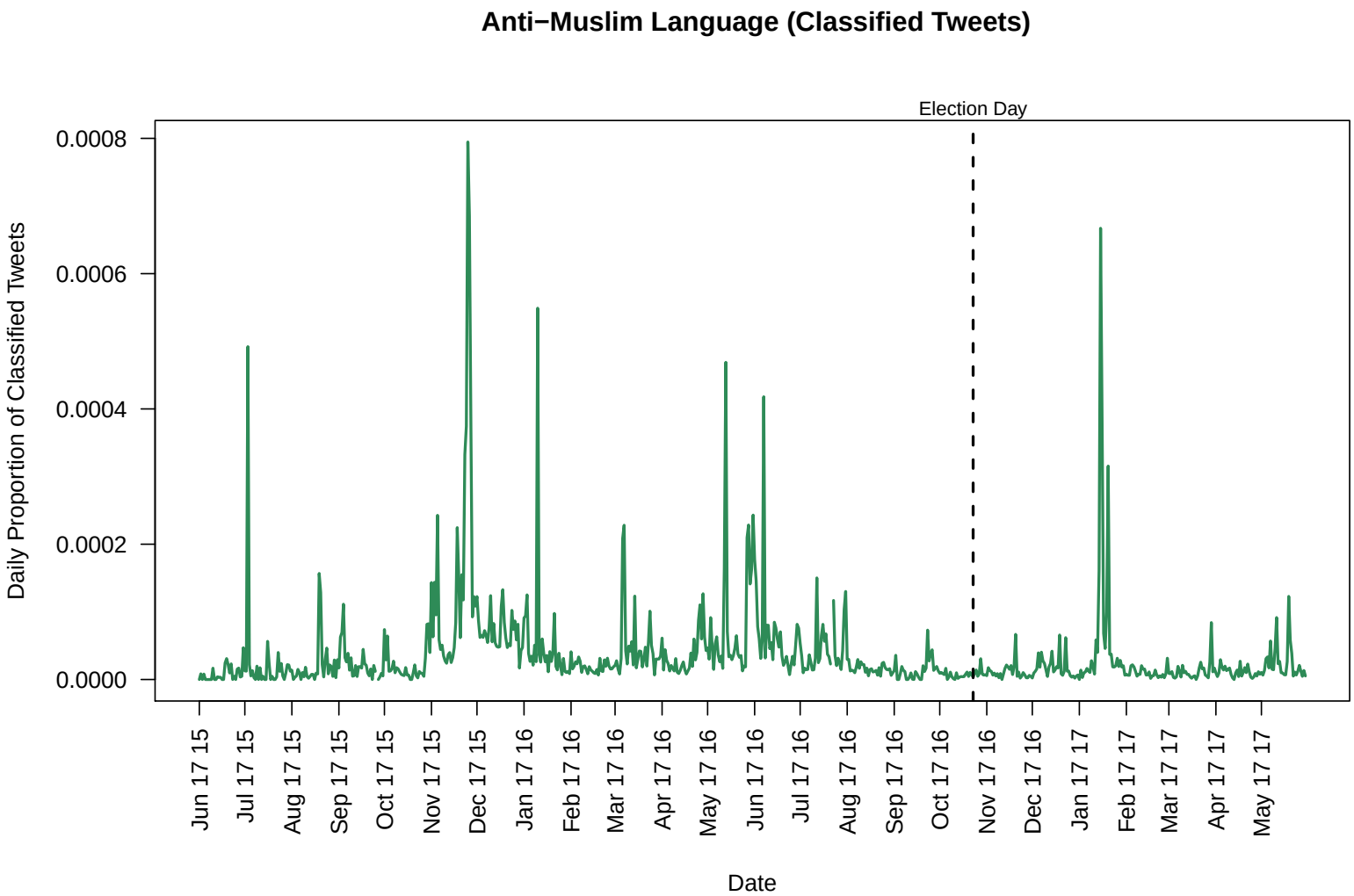
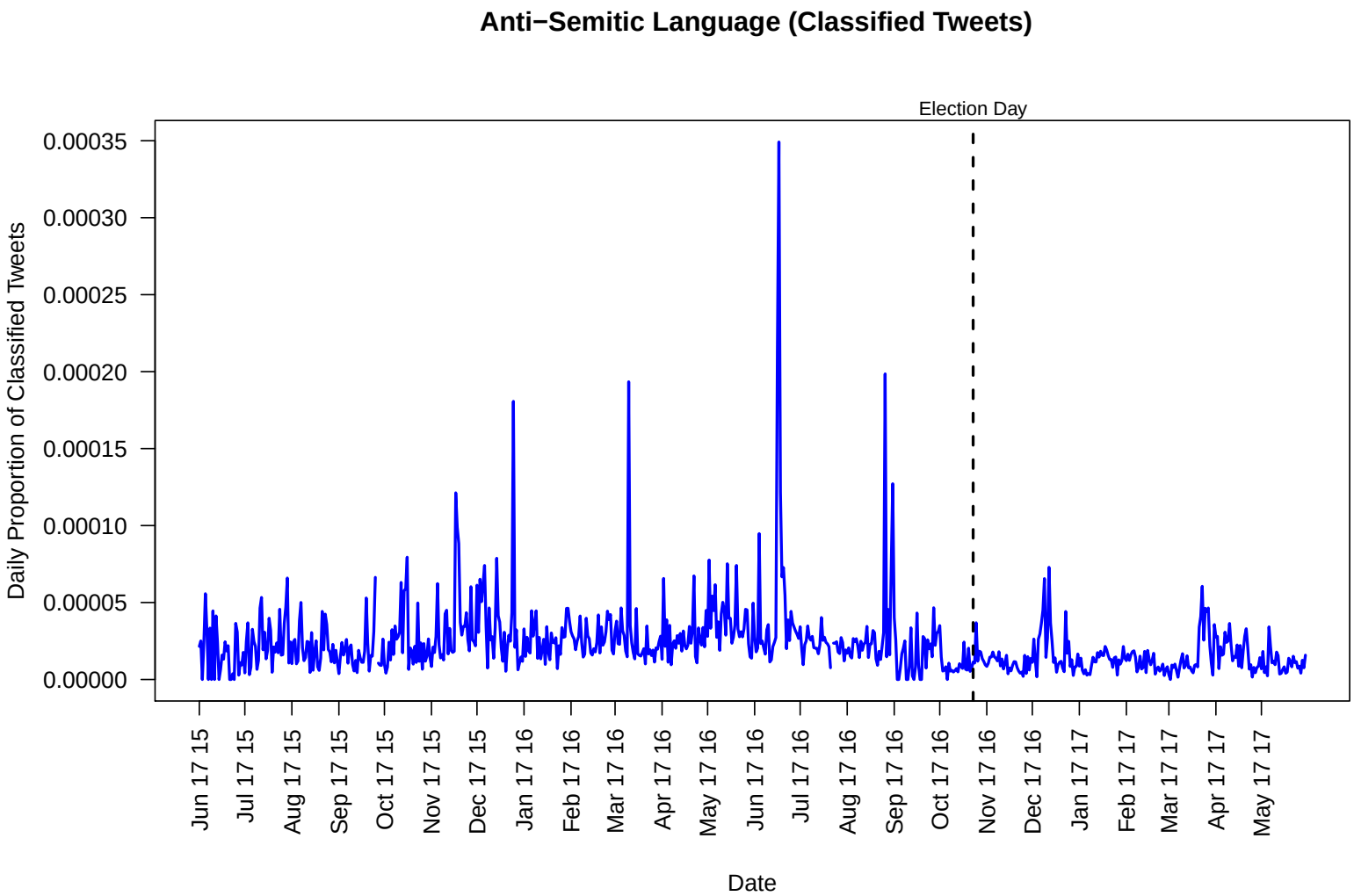
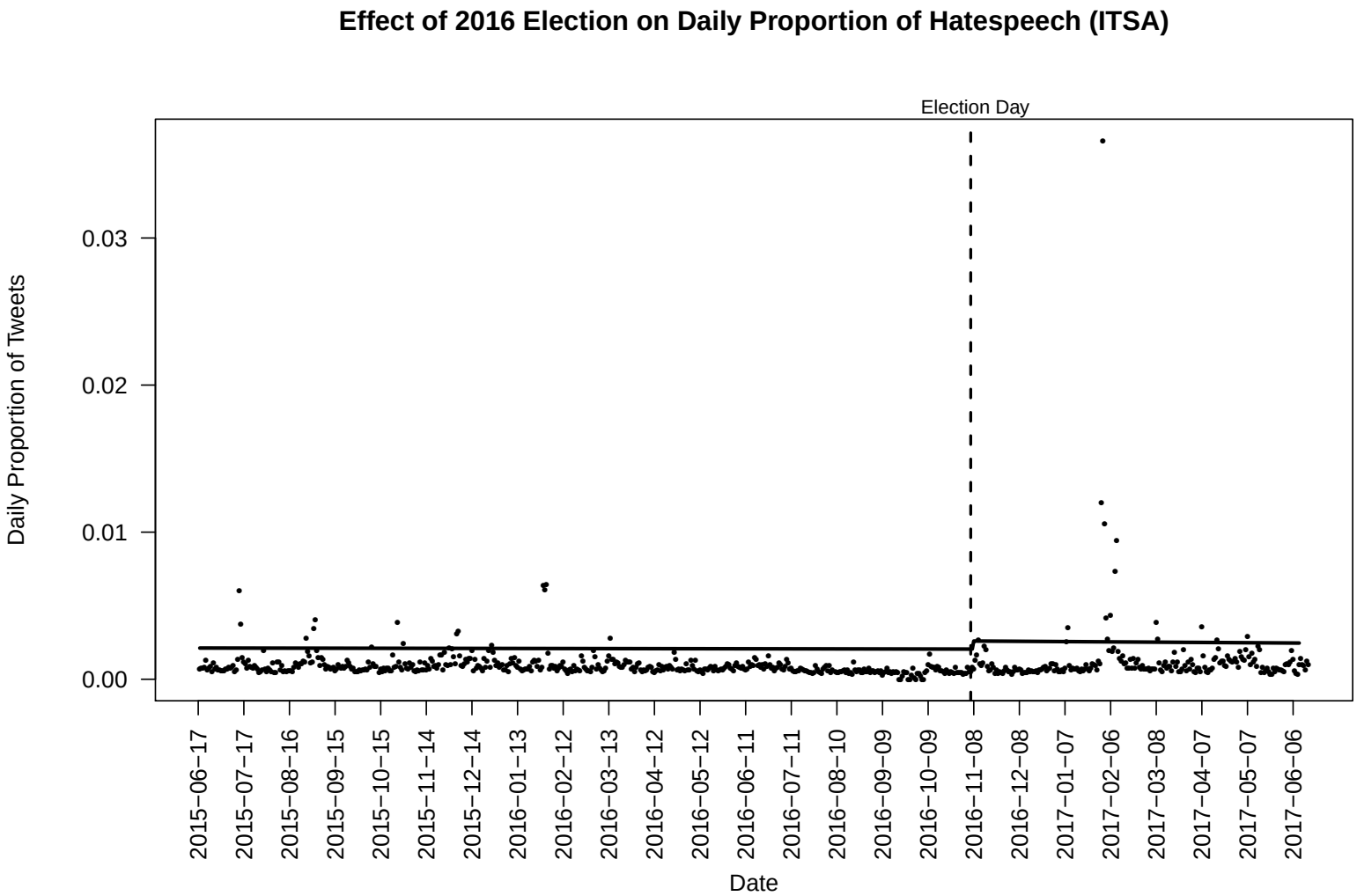


Figure : **Daily Proportion of Anti-Semitic Tweets in Trump Dataset**



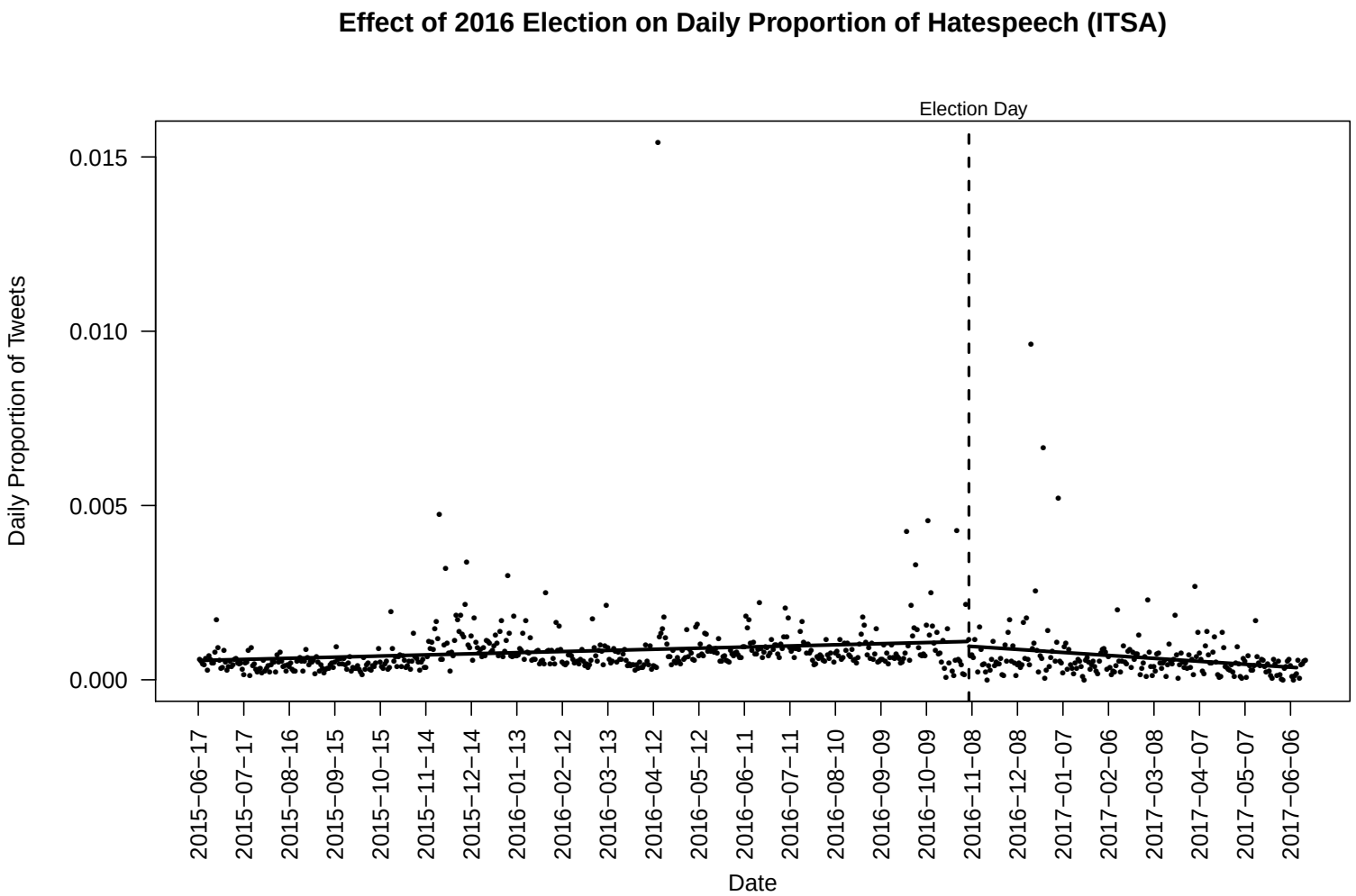
Interrupted Time Series Analysis - Hate Speech Trump Data Set

Figure : Effect of 2016 Election on Daily Proportion of Hate Speech Tweets (Trump Data)



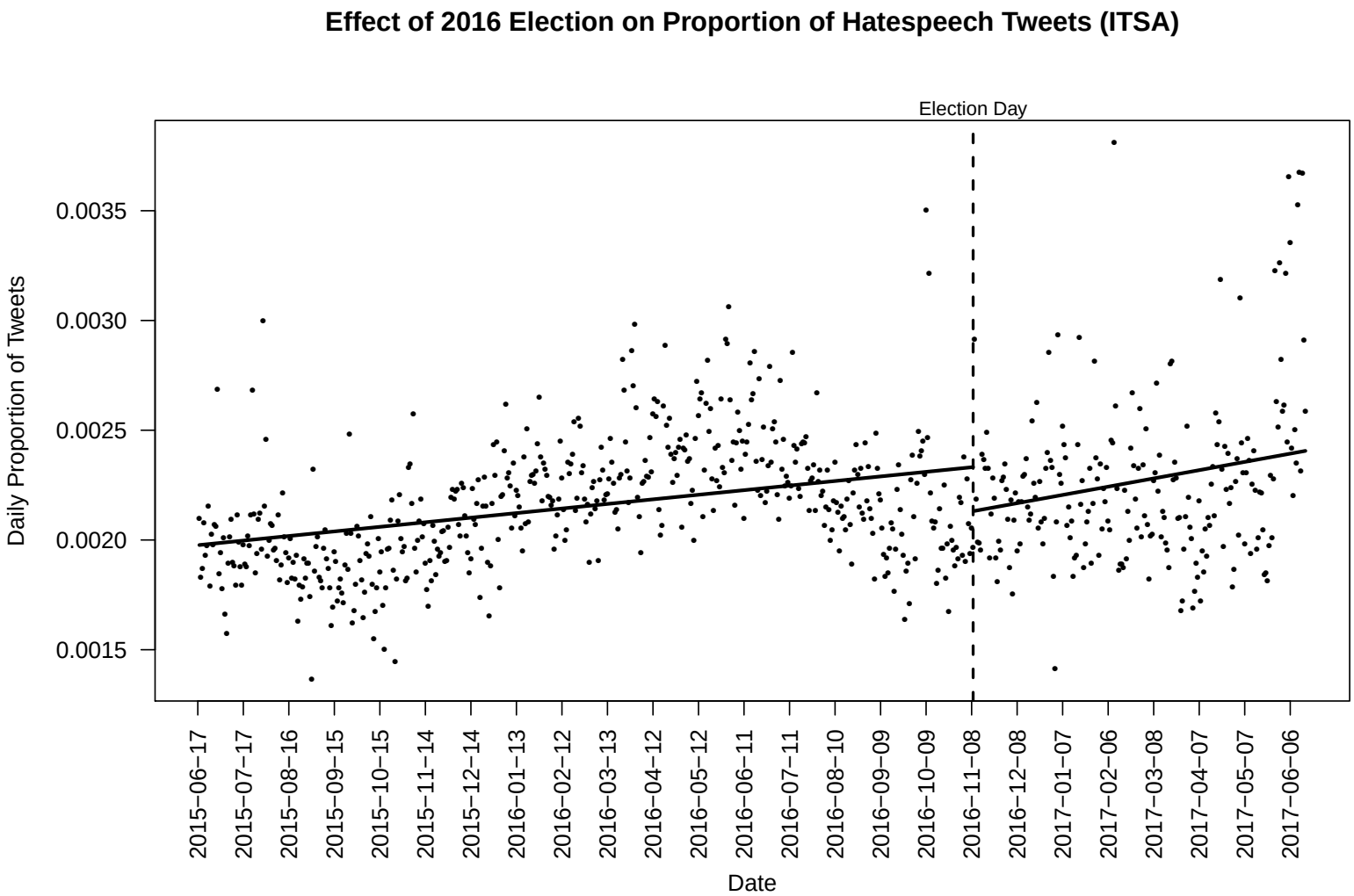
Interrupted Time Series Analysis - Hate Speech Clinton Data Set

Figure : Effect of 2016 Election on Daily Proportion of Hate Speech Tweets (Clinton Data)



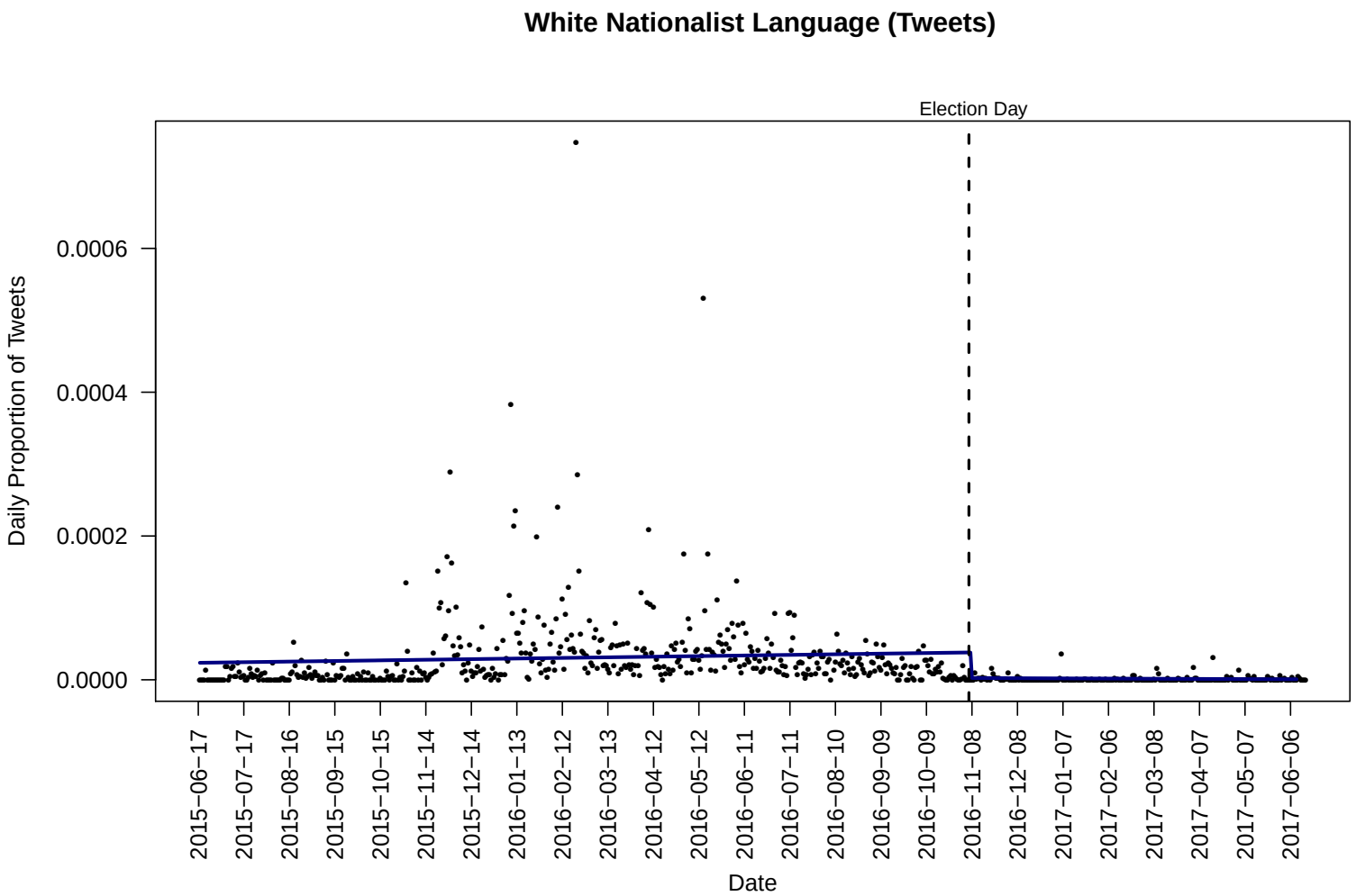
Interrupted Time Series Analysis – Hate Speech Random Sample

Figure : Effect of 2016 Election on Daily Proportion of Hate Speech Tweets (Random Sample)



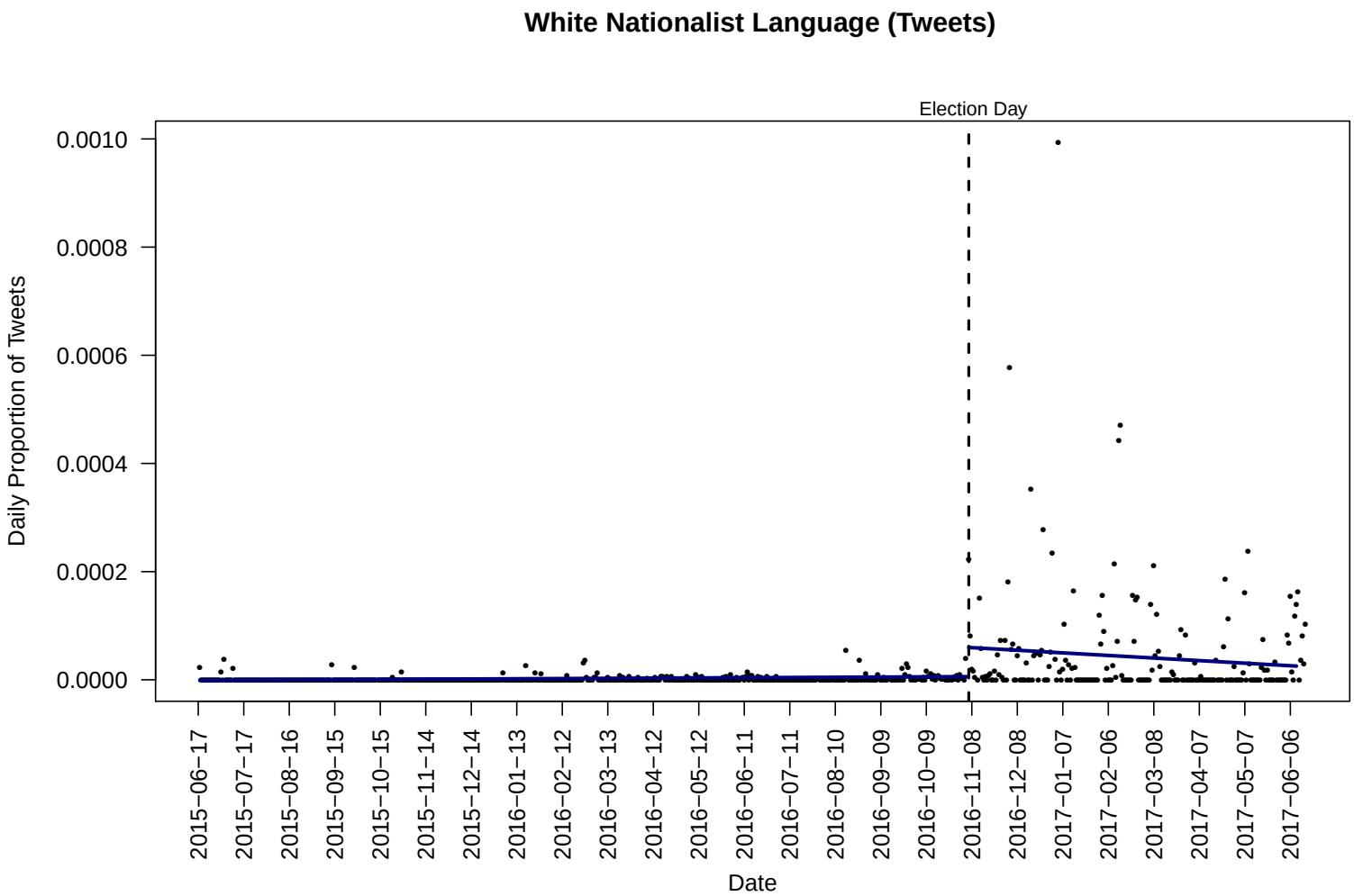
Interrupted Time Series Analysis - White Nationalist Trump Data Set

Figure : Effect of 2016 Election on Daily Proportion of White Nationalist Tweets (Trump Data)



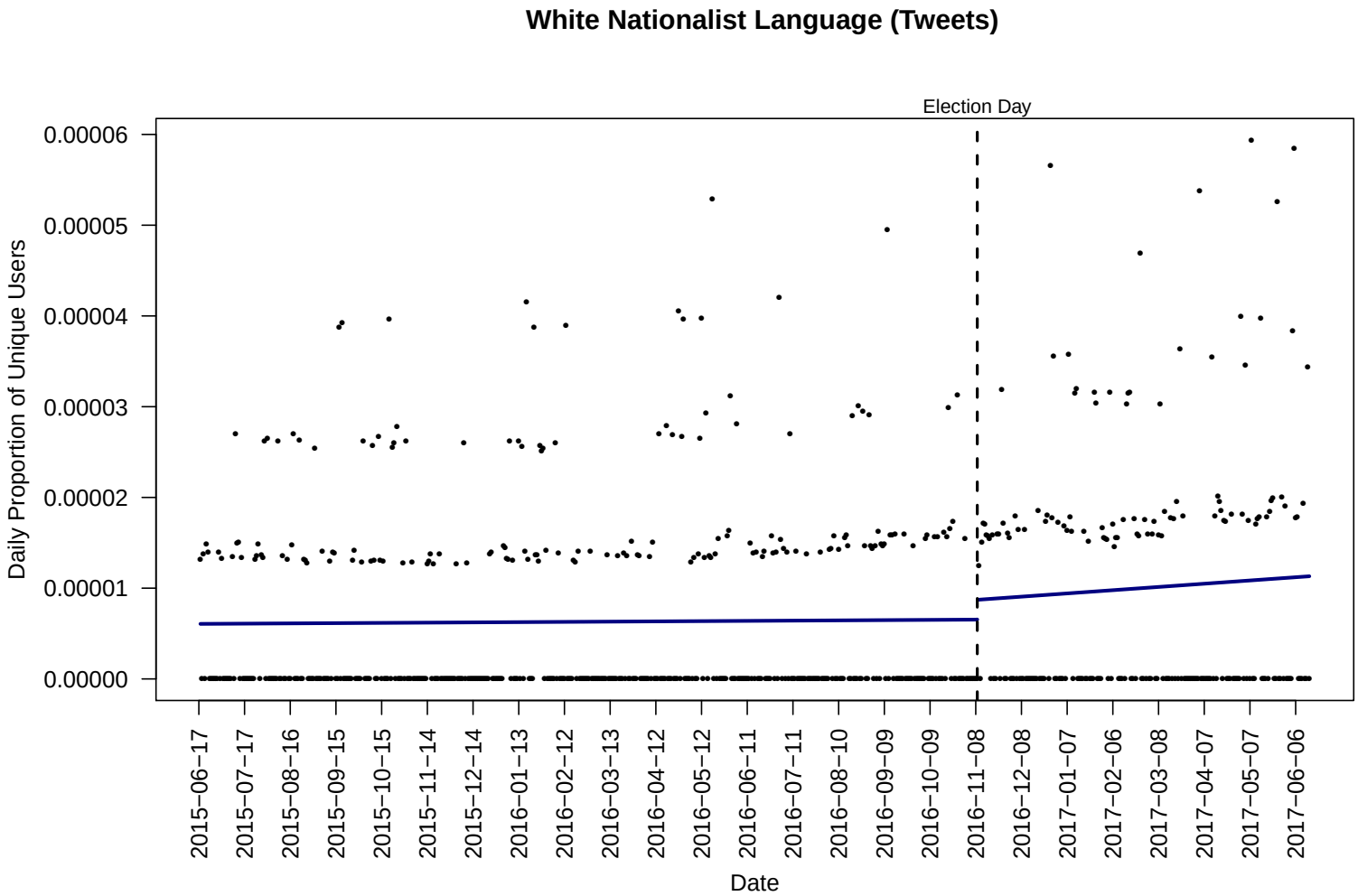
Interrupted Time Series Analysis - White Nationalist Clinton Data Set

Figure : Effect of 2016 Election on Daily Proportion of White Nationalist Tweets (Clinton Data)



Interrupted Time Series Analysis: White Nationalist Random Sample

Figure : Effect of 2016 Election on Daily Proportion of White Nationalist Tweets (Random Sample)



Conclusions from Machine Learning Augmented Dictionary Method

- Contrary to received wisdom, our analysis (thus far) does NOT show a systematic increase in hate speech over the course of the campaign or after the election.

Conclusions from Machine Learning Augmented Dictionary Method

- Contrary to received wisdom, our analysis (thus far) does NOT show a systematic increase in hate speech over the course of the campaign or after the election.
- Similarly, there are no significant increases in white-nationalist rhetoric over the course of the campaign in any of our three datasets.

Conclusions from Machine Learning Augmented Dictionary Method

- Contrary to received wisdom, our analysis (thus far) does NOT show a systematic increase in hate speech over the course of the campaign or after the election.
- Similarly, there are no significant increases in white-nationalist rhetoric over the course of the campaign in any of our three datasets.
- While there is a significant increase in white-nationalist rhetoric after Trump's election in the Clinton data—and a similar, though not significant—pattern in the random sample data, these effects are substantively very small and not present in our largest (Trump) dataset.

Robustness Test: Intuition

- Concern: are we missing other kinds of hate speech?

Robustness Test: Intuition

- Concern: are we missing other kinds of hate speech?
- Idea: Find a place with known hate speech, then compare daily tweets with that speech

Robustness Test: Intuition

- Concern: are we missing other kinds of hate speech?
- Idea: Find a place with known hate speech, then compare daily tweets with that speech
- Concept: Measure the semantic similarity between real world example of hate speech language and tweets by day

Measuring similarity

- **Doc2Vec** model was trained to generate vector representations of the text documents.

Measuring similarity

- **Doc2Vec** model was trained to generate vector representations of the text documents.
- Trains a neural network that is trying to predict words in the documents using the following data as an input:
  - Vectors of surrounding words
  - Individual document vectors
  - Subreddit vectors

Measuring similarity

- **Doc2Vec** model was trained to generate vector representations of the text documents.
- Trains a neural network that is trying to predict words in the documents using the following data as an input:
  - Vectors of surrounding words
  - Individual document vectors
  - Subreddit vectors
- This generates a vector embedding for each subreddit, which we can compare with the vector embedding for each day of tweets.

So what do we do?

- Measure daily semantic similarity between our three Twitter datasets and racist/misogynistic/ alt-right subreddits

So what do we do?

- Measure daily semantic similarity between our three Twitter datasets and racist/misogynistic/ alt-right subreddits
- See if the similarity increases over the course of the campaign or in the aftermath of the election (suggesting an increase in hateful language in each dataset)

Figure : Trump

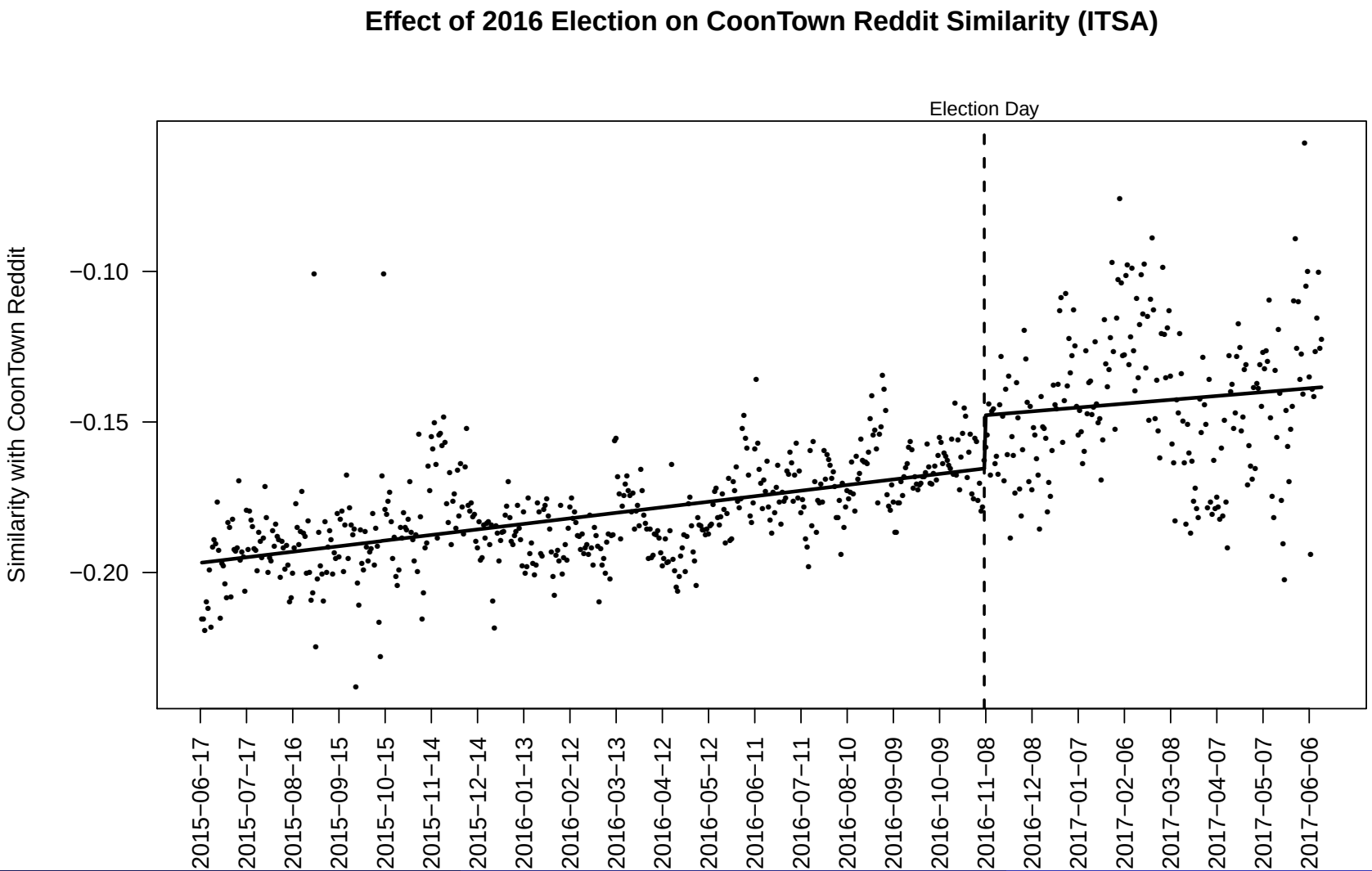


Figure : Clinton

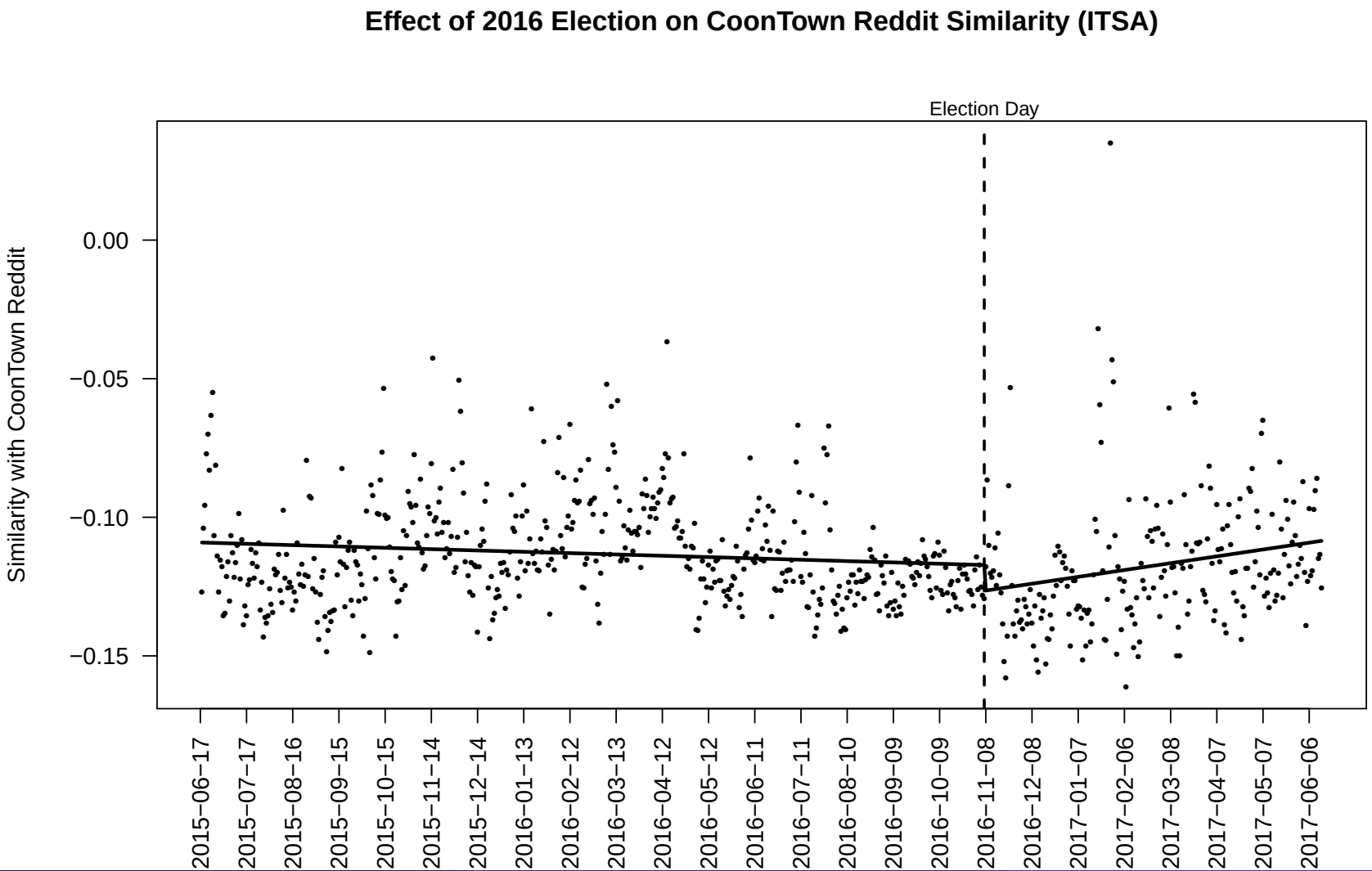


Figure : Random Sample

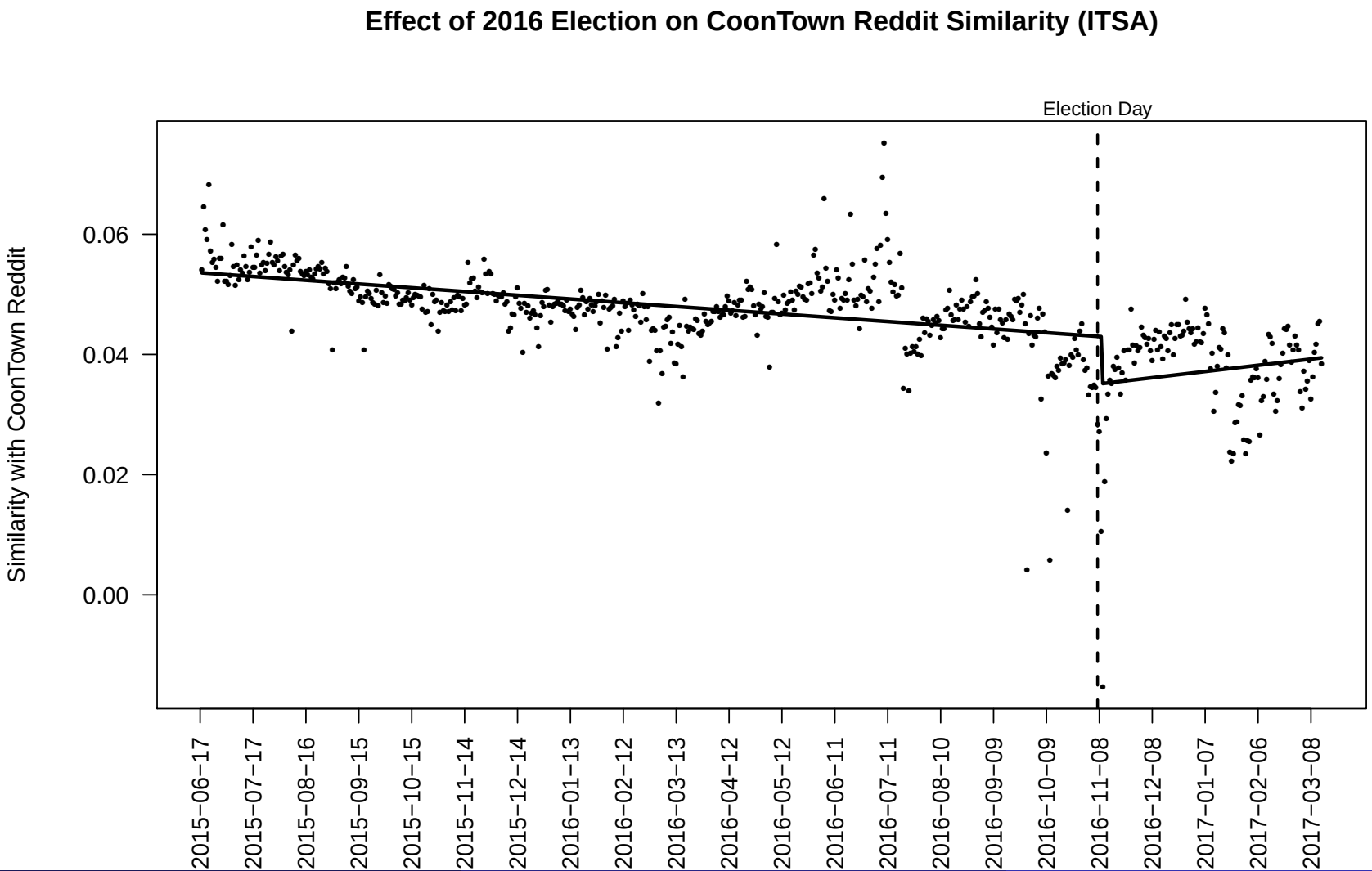


Figure : Trump

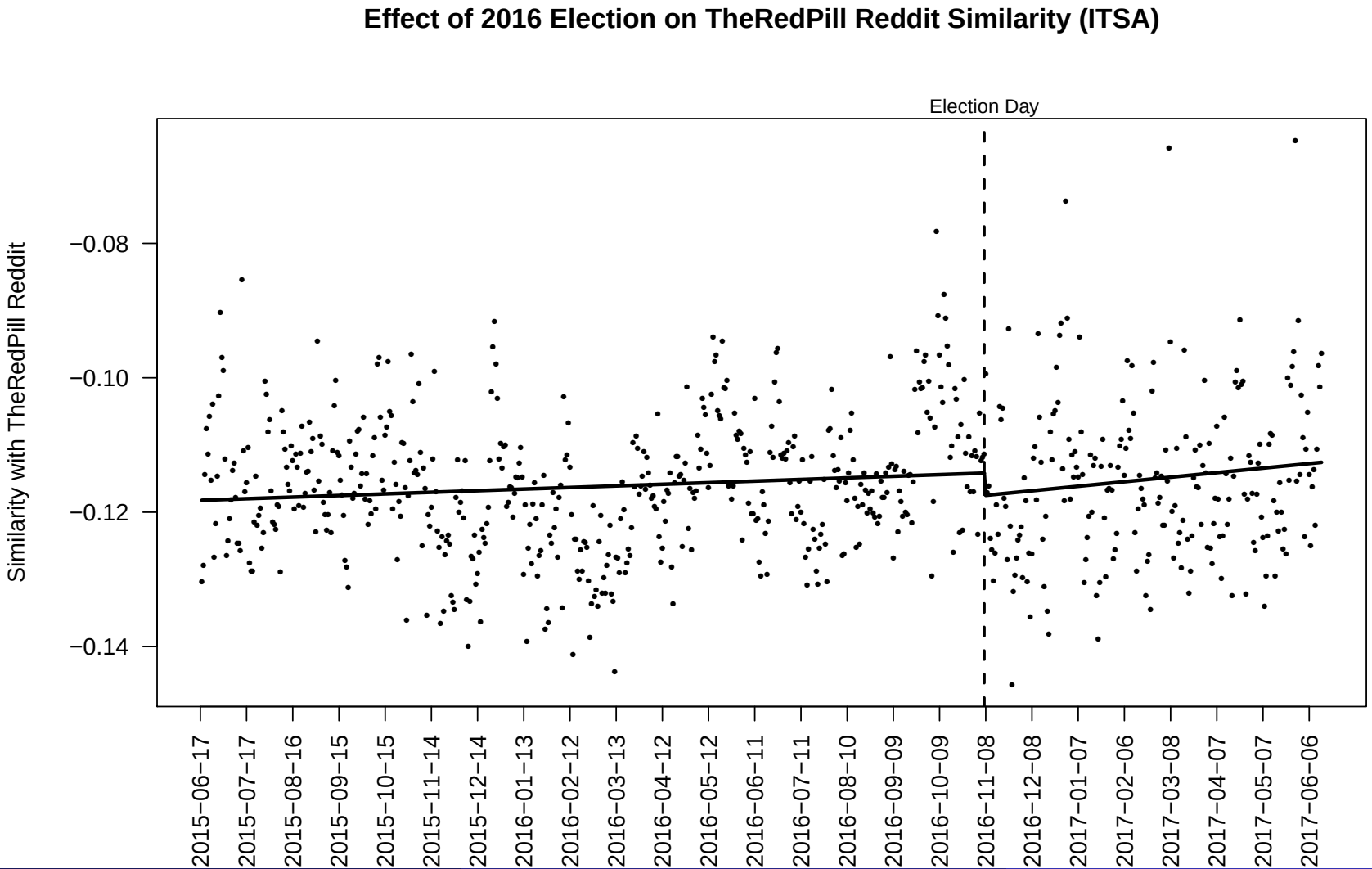


Figure : Clinton

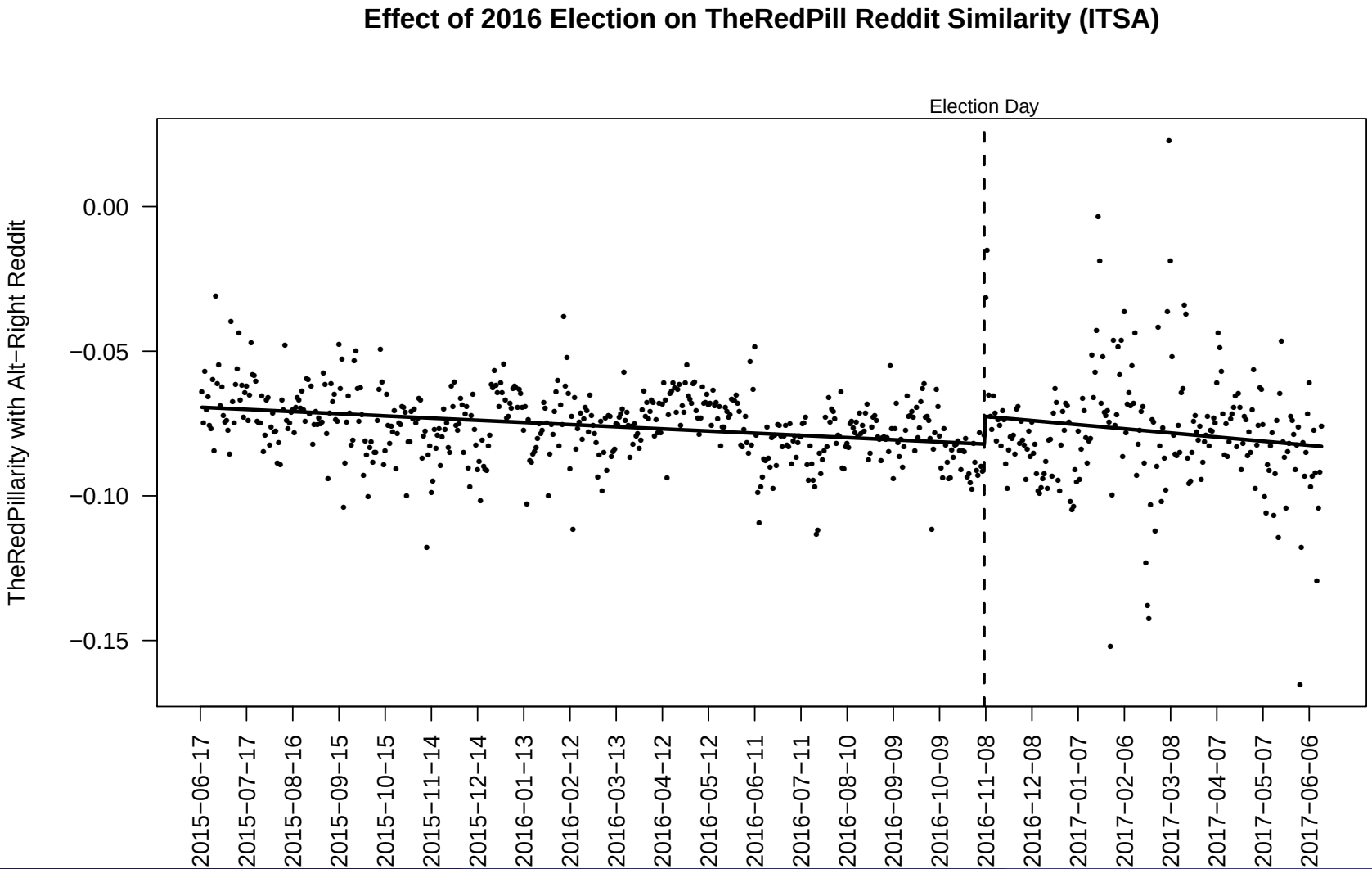


Figure : Random Sample

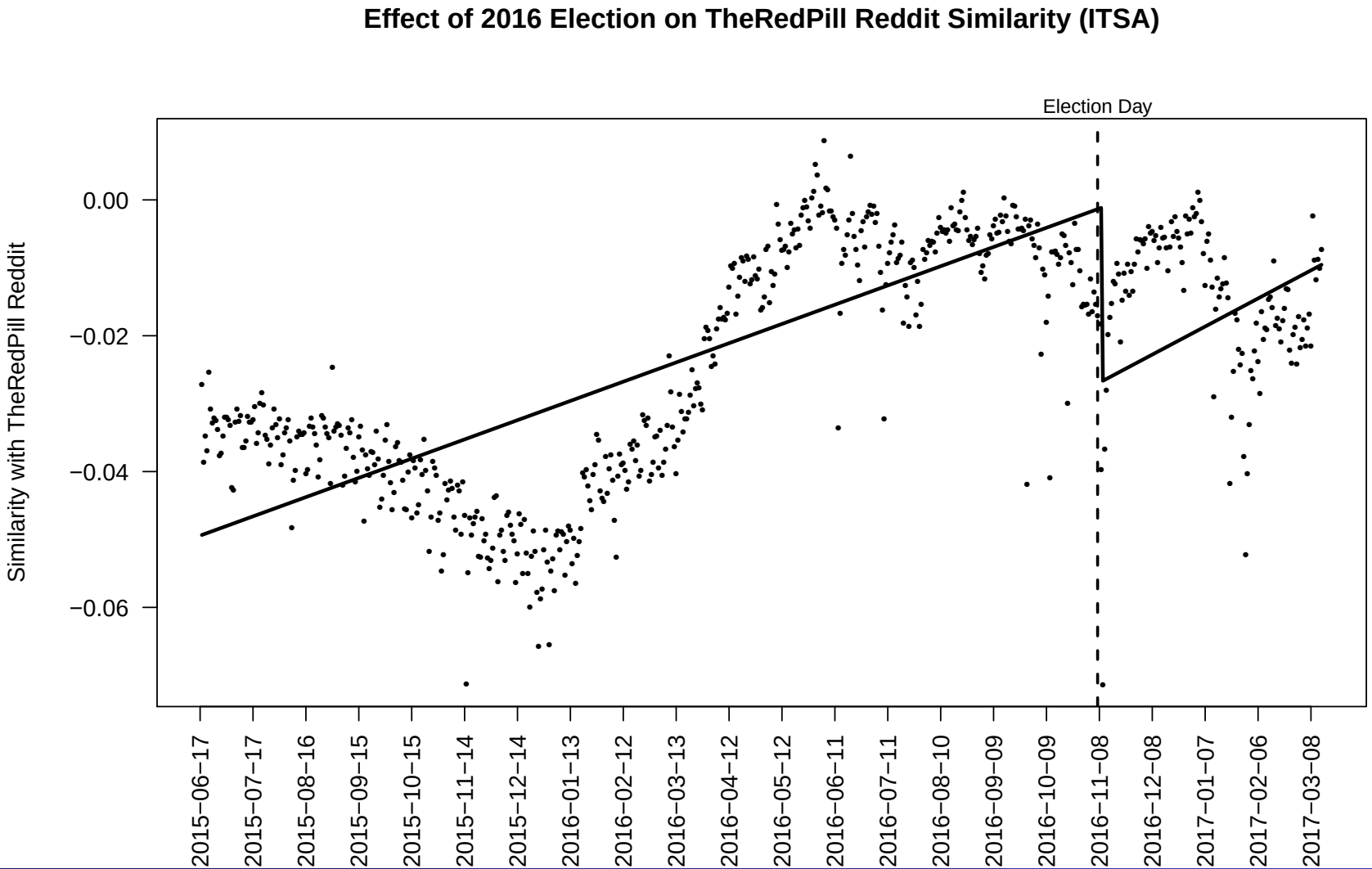


Figure : Trump

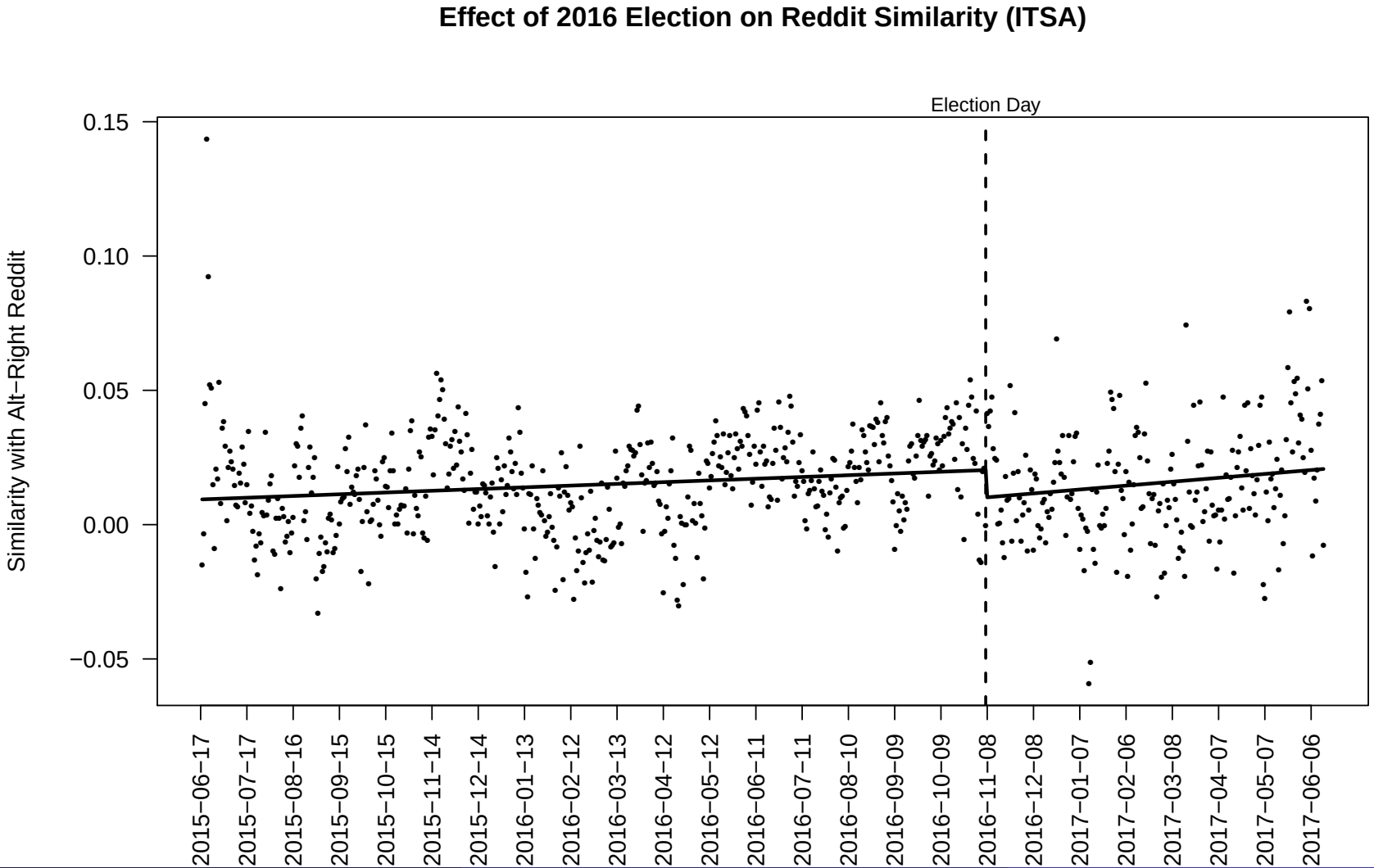


Figure : Clinton

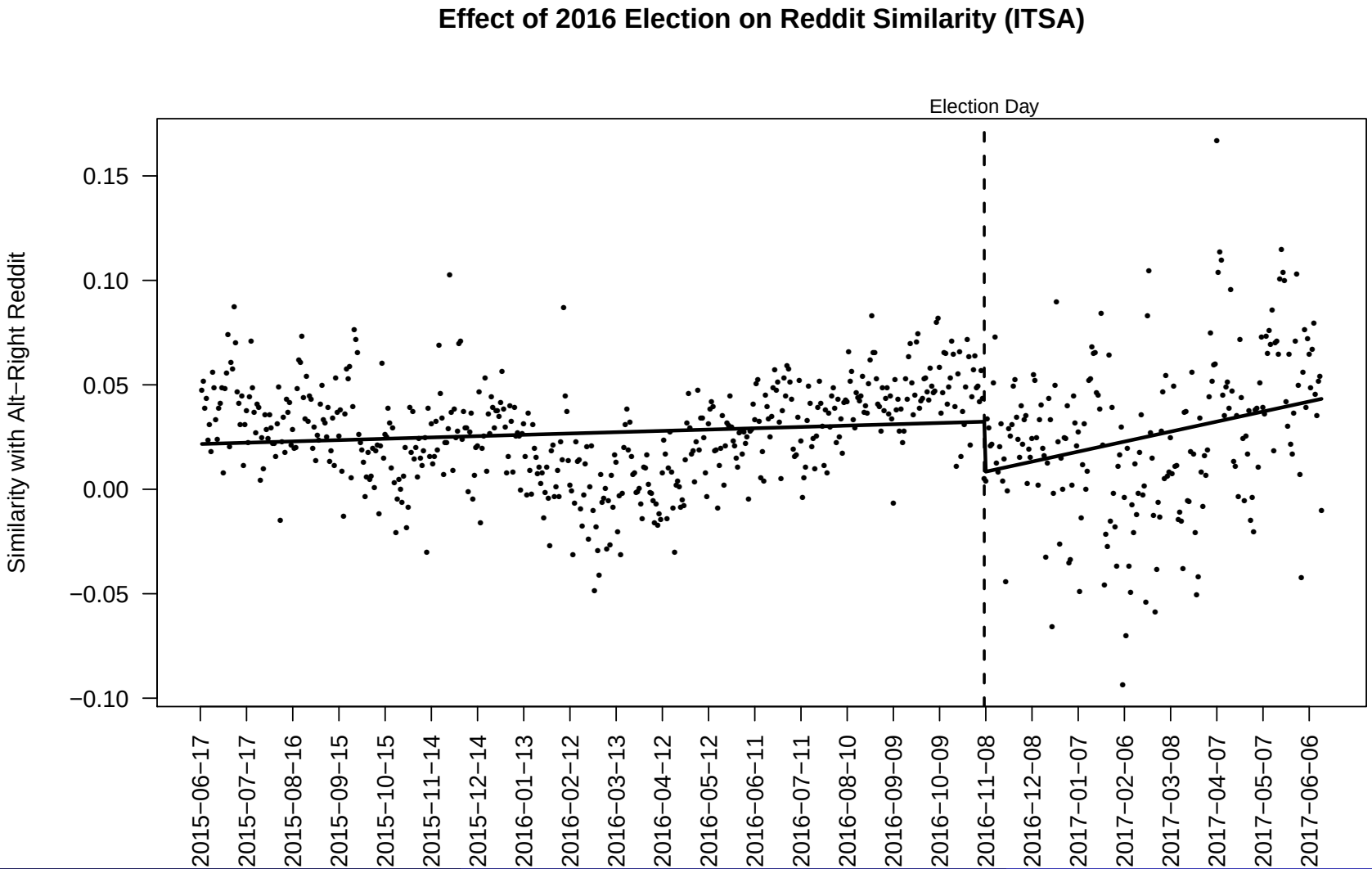
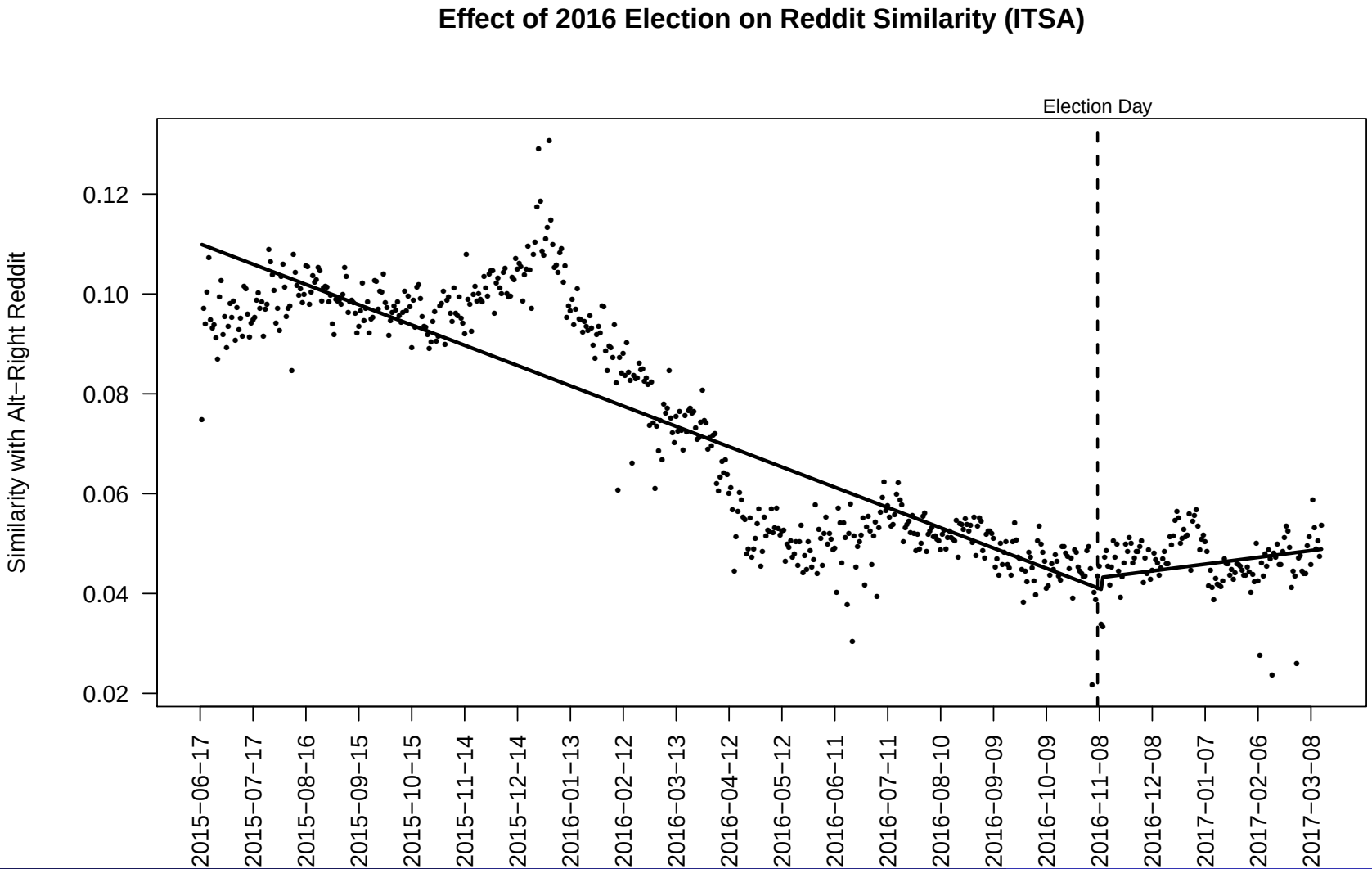


Figure : Random Sample



Conclusions

- Contrary to received wisdom, our results do NOT show a systematic increase in hate speech or white nationalist rhetoric during the election campaign using either dictionary or semantic similarity methods across our datasets.

Conclusions

- Contrary to received wisdom, our results do NOT show a systematic increase in hate speech or white nationalist rhetoric during the election campaign using either dictionary or semantic similarity methods across our datasets.
- There is some evidence of a small increase in white nationalist rhetoric after the election in the Clinton dataset, but these effects are very small and not consistent across our analyses.

Conclusions

- Contrary to received wisdom, our results do NOT show a systematic increase in hate speech or white nationalist rhetoric during the election campaign using either dictionary or semantic similarity methods across our datasets.
- There is some evidence of a small increase in white nationalist rhetoric after the election in the Clinton dataset, but these effects are very small and not consistent across our analyses.
- The primary divergence between our dictionary and semantic similarity methods occurs in the Trump dataset, where there is a significant increase in similarity to the racist subreddit over the course of the campaign and following Trump's election. We do not observe this pattern using dictionary-based methods.

The Caveats!

- Only a portion of Twitter
- Only Twitter, not other platforms (but question is "mainstreaming...")

The Caveats!

- Only a portion of Twitter
- Only Twitter, not other platforms (but question is "mainstreaming...")
- Nothing about attacks on individuals and their effects
- Nothing about attacks on journalists and their consequences

The Caveats!

- Only a portion of Twitter
- Only Twitter, not other platforms (but question is "mainstreaming...")
- Nothing about attacks on individuals and their effects
- Nothing about attacks on journalists and their consequences
- Nothing about offline hate-based attacks, threats

The Caveats!

- Only a portion of Twitter
- Only Twitter, not other platforms (but question is "mainstreaming...")
- Nothing about attacks on individuals and their effects
- Nothing about attacks on journalists and their consequences
- Nothing about offline hate-based attacks, threats
- Semantic similarity method is new and we have a lot to learn

The Caveats!

- Only a portion of Twitter
- Only Twitter, not other platforms (but question is "mainstreaming...")
- Nothing about attacks on individuals and their effects
- Nothing about attacks on journalists and their consequences
- Nothing about offline hate-based attacks, threats
- Semantic similarity method is new and we have a lot to learn
- So far, very little engagement with theory – just (an important!) test of received wisdom

Future Research

- Short term:
  - Robustness of semantic similarity method
  - Validation of subreddits
  - Including retweets in semantic similarity analysis

Future Research

- Short term:
  - Robustness of semantic similarity method
  - Validation of subreddits
  - Including retweets in semantic similarity analysis
- Long term:
  - The effects of other campaign and external events on the popularity of hate speech.
  - Hate speech by geographic region / effect of local events
  - Trends on other online platforms.
  - The connection between online hate speech and offline hate crimes and bias incidents.
  - Alternative sources of text (Gab? Voat? 8chan?)

Thank you!

Thank you!

Subreddits used in our analysis

Subreddits					
Alt-right and hate	**Political**	**Gender issues**	**African American culture**	**Sport**	**Entertainment**
AntiPOZi	EnoughTrumpSpam	GenderCritical	BlackPeopleTwitter	baseball	Android
CoonTown	HillaryForPrison	TwoXChromosomes	blackladies	soccer	StarWars
DebateAltRight	PoliticalDiscussion	lgbt	hiphopheads	sport	books
PussyPass	Republican				gaming
WhiteRights	The_Donald				movies
european	democrats				
sjwhate	hillaryclinton				
TheRedPill	politics				
MensRights					

Validation of Method I: Similar subreddits are located in same space