Words of Warning

Inferring intent from textual evidence via machine learning.

The Insurrection of January 6th, 2021

Previous Efforts

“(A Data Scientist at Twitter) argued that, on a technical level, content from Republican politicians could get swept up by algorithms aggressively removing white supremacist material. Banning politicians wouldn’t be accepted by society as a trade-off for flagging all of the white supremacist propaganda...

... the company told Motherboard that this “is not [an] accurate characterization of our policies or enforcement—on any level.””

https://www.vice.com/en/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too

MOTHERBOARD
TECH BY VICE

Why Won’t Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too.

A Twitter employee who works on machine learning believes that a proactive, algorithmic solution to white supremacy would also catch Republican politicians.

By Joseph Cox By Jason Koebler

April 25, 2019, 12:21pm Share Tweet Snap

At a Twitter all-hands meeting on March 22, an employee asked a blunt question: Twitter has largely eradicated Islamic State propaganda off its platform. Why can’t it do the same for white supremacist content?

Donald Trump banned from Twitter

Criticized by some as “too little, too late”

Project Proposal

1. A new toolkit for content review and flagging, allowing social networks such as Twitter to be more confident in their moderation.
2. Model built for inference, not strictly prediction
3. Designed to be used in conjunction with human review
4. High interpretability
5. Engineered features using topic modeling and a synthesis of word vectors and sentiment analysis.
    a. Identify topics in light of particular legal precedent
    b. “Riot Index” - the intersection of topic and sentiment.

Legal Background

1. “Imminent Lawless Action” - *Brandenburg vs. Ohio*
2. “Incitement” - “inchoate” offense, dependent strictly on intent.
3. “Fighting Words”: Decision in *Chaplinsky vs. New Hampshire* asserted that:

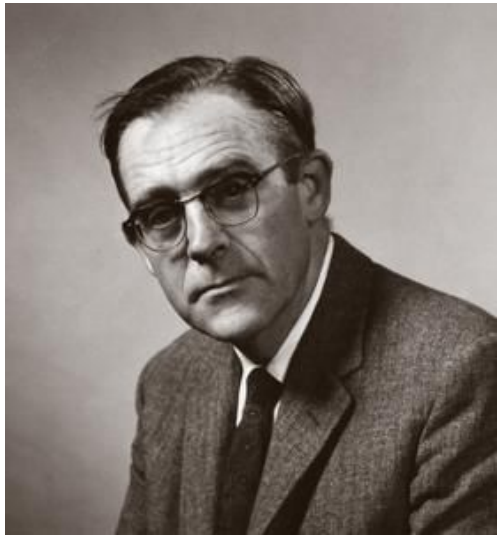
> "insulting or 'fighting words', those that by their very utterance inflict injury or tend to incite an immediate breach of the peace" are among the "well-defined and narrowly limited classes of speech the prevention and punishment of [which] ... have never been thought to raise any constitutional problem."

Symbols of American Extremists

One of the chief ways we recognize white supremacists, conspiracy theorists and the like — and how they recognize each other — is through the appropriation and use of symbols.

Images from: https://www.cnn.com/2021/01/09/us/capitol-hill-insurrection-extremist-flags-soh/index.html

What is coded language?



- Also called a Dog Whistle, Loaded, or ‘High-Inference' language
- Analytic and ethical philosopher Charles Stevenson provides a useful definition:
  - Words that carry more meaning than a simple description.
  - They are not neutral, but have a “magnetic” effect, bound to particular value judgements.
  - Attempt to elicit an instinctive emotional response, often to override reason.
- Carry “emotional valence” - presupposing a value judgement intended to elicit *prima facie* emotional response leading to action.

Some known predictors

“The Storm”

“Deep State”

“Drop the hammer”

“Battle stations”

“Stop the Steal”

https://www.brandeis.edu/now/2021/january/trump-language-capitol-riot-mcintosh.html

White Supremacist Dog Whistles

Most Predictive (Positive set):

- “HH” or “88” for Heil Hitler: rendered as 🎱🎱 on Twitter to evade bans.
- “SS” for the Schutzstaffel, rendered ⚡⚡.
    - Fortunately for our model, both of these usages are actually *more* definitively identifiable.

Hit/Miss (validation set):

- Anything Qanon - less commonly discussed among non-adherents prior to Jan 6
- “1933” or “frens” - specifically used by neo-nazis but could show up elsewhere easily
- Mention of Oathkeepers, 3%, Proud Boys, Boogaloo, Civil War. Moderately predictive.

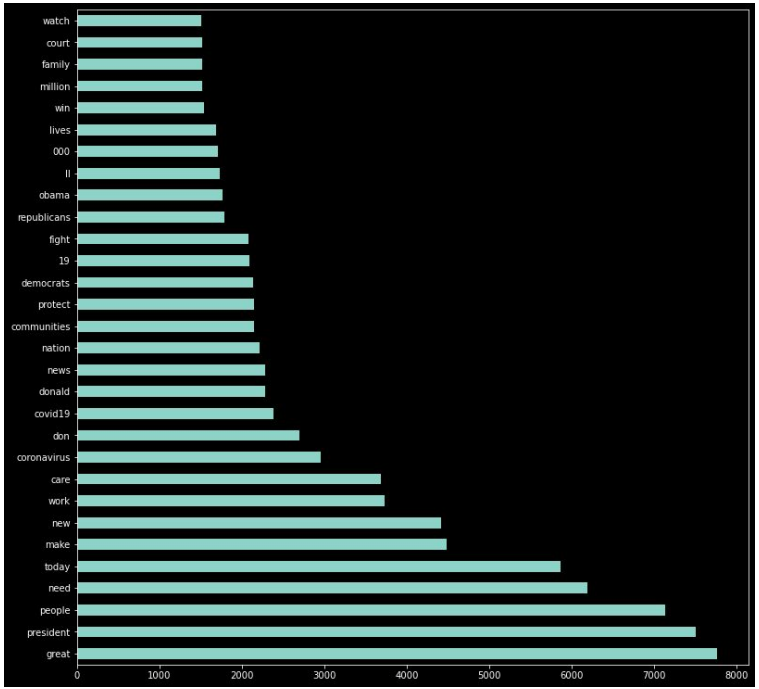
Methodology

1. Create a positive and negative case.
    a. Trump (actually banned) vs. other politicians
    b. Definitive topics (Nazi codes) vs. random tweets from same timeframe
2. Create a separate validation set to examine
    a. Politicians implicated but not confirmed
    b. Topics that may include both discussion of and actual adherents (Qanon, Proud Boys, etc.)
3. Continually update model based on critical analysis of results.

Model Selection

1. Logistic Regression
    a. Examine coefficients for interpretation and implementation in feature engineering.
2. Random Forest
    a. Extract individual decision trees for analysis
3. RNN
    a. Known to be useful for NLP tasks
    b. Attempt to achieve highest “predictive power” - see what gets flagged by a black box version of the model

Initial Findings



Future Plans

- Continual improvement of models with new data
  - Distinguish ‘evergreen’ vs. trending keywords & ngrams
  - Apply Timeseries modeling to data, week by week, from 2020 election to Jan 6th
- Web app allowing direct text input or twitter handle
- Apply model to Gab and Parler “leak” data
- RNN to achieve maximum predictive power.
- Categorize keywords and ngrams from initial models, feed into:
- Unsupervised learning on both original data and new iterations to better understand model and predictors.
- Engineer ‘Riot Index’