

Brexit and healthcare were the most consistent and persistent topics. Women's rights, Trump, and climate also appeared. Immigration, religion, economy did not.

Twitter data acquisition workflow

11 Million "hydrated" tweet ids

"Re-hydrate"

Clean and Prep

3.1 Million tweets for modeling

- Acquired from "The Twitter Parliamentarian Database"
- Over 48 hrs. to download
- May 2017 to June 2019
- Primarily EU politicians but includes US, Australia, and Canada

- Twitter developer account! Load into Pandas
- Tools: Hydrator, Twarc
- Over 48 hrs. to download and convert to csv

- Extract tweets in English
- Regex to strip numbers, punctuation, symbols, URL
- Delete 'text' field duplicates, nan

Exploratory modeling: is it an elephant, a giraffe, or a T-rex? The surprise was what did not appear as a topic: immigration and religion

Preliminary modeling with LSA & NMF Stemming and n-grams did not help

NMF 55k words, 75 topics A few topics that made sense

- TFIDFVectorizer and CountVectorizer min_df = 0 resulted in 2.9 Million features
- Experimented with various topic counts and min_df

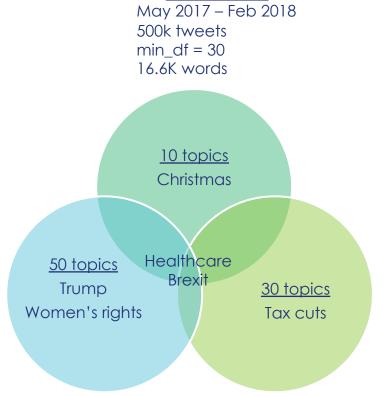
- Snowballer stemmer topic output was gibberish
- N-grams caused feature space to explode
- Min_df = 30
- Added my own stop words to scikit library
- Model ran for nearly 4hrs.

- Healthcare
- Trump
- Public education
- Brexit (3 topics)
- Women's rights

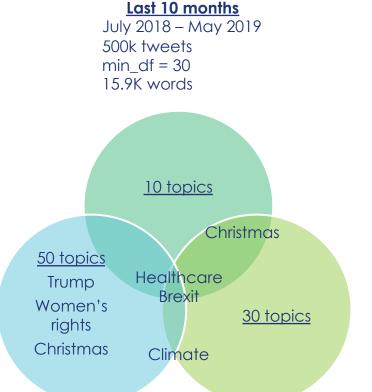
NMF with different time periods and n_component hyper_parameters Healthcare and Brexit consistently appear as topics

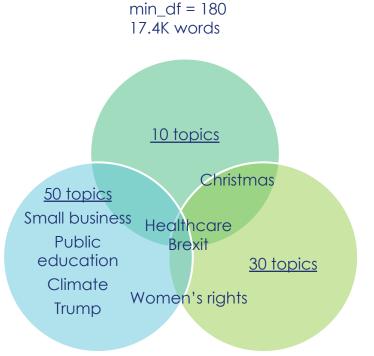
NMF different time periods and n_component = 10, 30, and 50

(stop_words = 'English' union my custom stop words)



First 8 months





All months

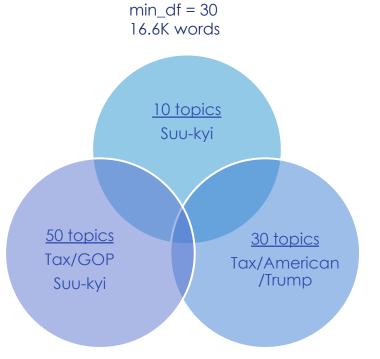
May 2017- June 2019

3.1M tweets

Corex was less interpretable Brexit and Trump dominant topics for last 10 months and all month time periods

Corex unsupervised different time periods and n_component = 10, 30, and 50

(stop_words = 'English' union my custom stop words)

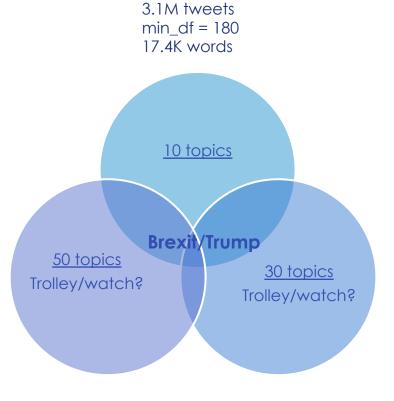


First 8 months

May 2017 - Feb 2018

500k tweets

Last 10 months July 2018 - May 2019 500k tweets min df = 3015.9K words 10 topics Trumpcare Jewish Holidays 30 topics Brexit 50 topics Boko-Haram abductions Boko-Haram abductions Ukraine elections



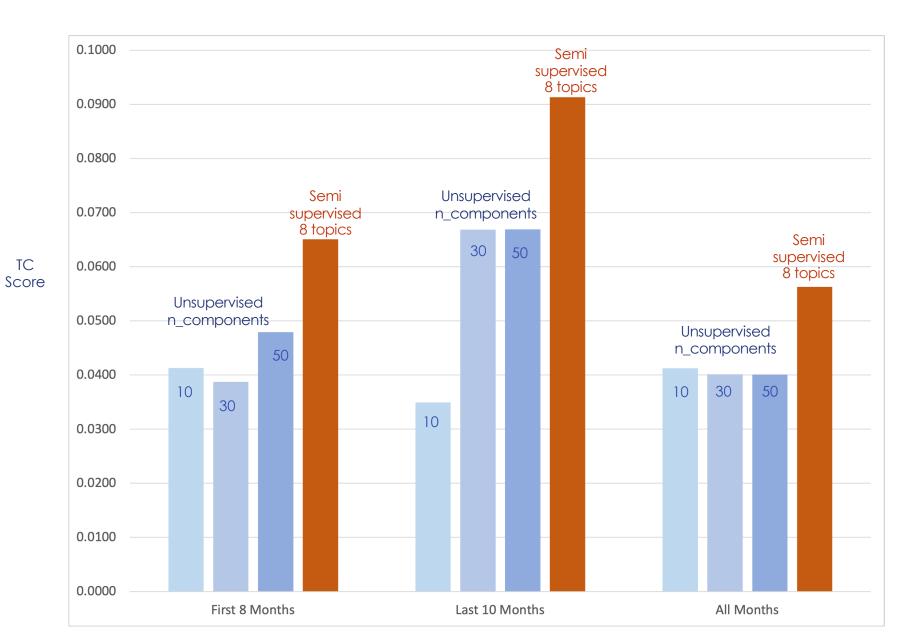
All months

May 2017- June 2019

Topics difficult to interpret and diffuse

Topics more interpretable and defined

Corex semi-supervised resulted in higher TC scores

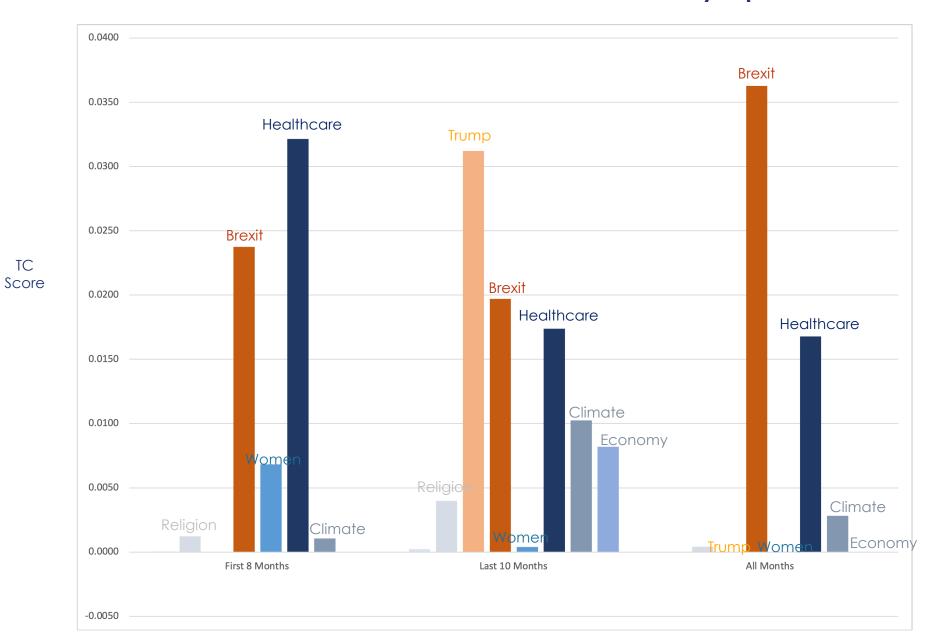


Semi-supervised topic anchor words

immigration, immigrant
Christian, Hindu, Muslim, Jewish, religion
trump
brexit, eu, leave
women, violence, rights
health, care, Insurance, mental
climate, change
jobs, economy, business

TC

Brexit and healthcare consistent key topics



- Immigration and religion absent
- Last 10-month period sharper topic definition
- Is Brexit subsuming topics like economy?

Finally used spacey and ScatterText to map overall word frequency and relationship on Liberal v Conservative axes

