What are politicians tweeting, anything meaningful?

By: Alex Ezazi

Abstract

This was a topic analysis of tweets by mostly EU politicians over a three period from May 2017 to June 2019. The goal was to determine any dominant topics and any changes over time.

Design

The data was in the form of 'hydrated' tweets from "The Twitter Parliamentarian Database" (https://figshare.com/articles/dataset/The_Twitter_Parliamentarian_Database/10120685).

The design approach was to analyze data with LSA, NMF, Corex, and ScatterText to identify topics. Time periods of the first 8 months (500k tweets), the last 10 months (500k tweets), and all months (3.1M tweets) were analyzed.

Data

The original database was over 11 million tweets in the form of 'hydrated' tweet ids. A twitter developer account was required to download the data. The tweet id database was 're-hydrated', first to a jsonl file, and then to csv (tools: Hydrator, Twarc) for loading into Pandas.

The English tweets were extracted, 'text' field duplicates and nan were deleted, regex was used to strip punctuation, symbols, numbers, URL, and to convert to lowercase. A total of 3.1 million tweets remained for modeling.

Model

After exploratory modeling with LSA, NMF, and Corex, a systematic analysis over the defined time periods was performed using NMF, Corex unsupervised, and Corex semi-supervised using anchor words gleaned from the unsupervised models and words related to immigration and religion as topics.

TFIDFVectorizer was used for all modeling.

The unsupervised models were run with n_components = 10, 20, 30, 40, 50 for each time period. The semi supervised Corex model was run with anchor words corresponding to 8 topics for each time period.

For the first 8-month (500k tweets) and last 10-month periods (500k tweets), min_df = 30 was used resulting in approximately 16k words. For the all months period (3.1M tweets), min_df = 180 was used resulting in approximately 17k words.

A separate word frequency analysis and visualization was performed using Spacey and ScatterText. Code for ScatterText was taken from a tutorial.

Tools

- Data acquisition: Twitter developer account, Twarc, Hydrator
- EDA and Cleaning: Pandas, numpy
- Topic Modeling: LSA, NMF, Corex unsupervised, Corex semi-supervised
- Word frequency and visualization: Spacey, ScatterText
- Presentation: Excel, PowerPoint