DDOG : IMPEACHMENT DAY TWEET ANALYSIS

Introduction

Our goal in this analysis is to determine whether people are for or against the impeachment based on tweets during the impeachment hearing. Furthermore we want to map this orientation per state. Political discourse on social media is becoming common practice, especially around election time.

Getting Data

We collect the tweets from the hashtag about the impeachment for several days. In this analysis we focused on the tweets on November 15th on #ImpeachmentHearing (142,106 tweets) and #ImpeachingHearing (55,000 tweets).

To do so we used the twitter api and the module tweepy on python. The api allows you with a free developer account to get at most the last 100 tweets every 5 seconds. This rate is big enough to handle most situations thus we manage to collect nearly all tweets. We coded an auto-regulatory time between requests based on the number of tweets collected in the previous request. We set up a raspberry pi zero to act as a server. Now from anywhere we are able to collect tweets about any topics.

Our dataset is composed of 200,000 tweets (with possible overlap) and over 30 columns for each tweet : id, the tweet itself, the location, information about the user (id, location..)

Data Cleaning and extraction of a location

We first clean the text of the tweets using regex, we remove urls and retweet mention. As mentioned before, we wanted to get the location of the tweets, but less than 1% do have one. Thus we extracted the location from the user's profile. This is a user based input thus can be false and is not formatted but it's the best we can get out of our data. 'New Jersey, USA', 'Where else? Georgia', 'Palm Beach, FL', 'Vancouver, BC', 'Earth citizen'.

We tried to use geocode api to get coordinates, but those api where slow and expensive. We headed towards a simpler implementation. We tokenize the text and find most of the states, except "New York" or "N.J.". We used bigrams to complete our state finder, some issues remain with Washington DC and the state Washington.

State classifier analysis : we classified 62% of tweets which user has indicate a location (41% on the whole dataset). The remaining part is mostly names of city, foreign countries, vague or non sens location. After going manually through hundreds locations classified we did find a classification error rate under 4% mostly Washington and Indiana (preposition 'in') and thus we can be pretty confident in our classifier.

We end up with a dataset of over 140,000 tweets with a cleaned text and a state.

Sentiment Analysis

To understand on which side people are, we performed sentiment analysis using the compound score (denoted C here) from VADER. As I can tweet something positive for the impeachment or supporting trump against impeachment, we need a second classifier. We first used named entity analysis, however this was extremely computationally expensive as run our analysis on 100 tweets took 2min. We created our own classifier, it acts as a counter +/- 1 for each pro/anti words, we end up with a Vocabulary score : V

		Handmade classification on keywords	
		Pro-impeachment words	Anti-impeachment words
Vader analysis	Positive sentence		
	Negative sentence		

Using our two classifiers, we define the weighted compound sign(V)*C. This method runs in a few minutes. This metric grasp the real user's intention and side

One user = On vote

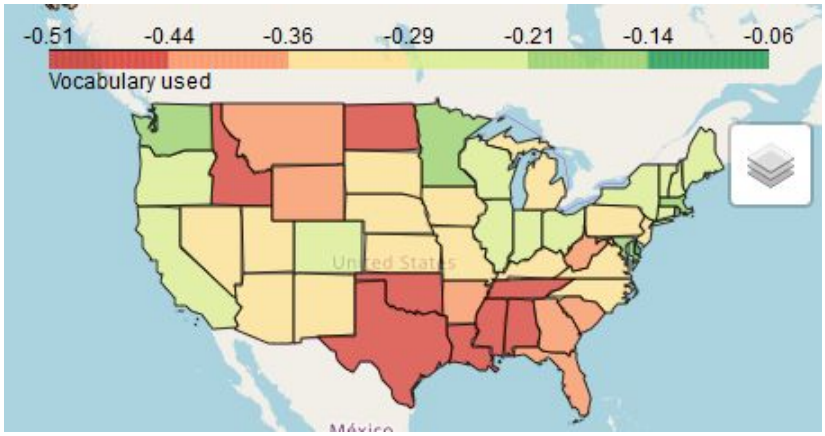
To avoid bias of user tweeting dozens of times, we group our data per user. We want to classify people, thus we are selecting the most extreme position of the user. The V score will be the most extreme V score of it's tweets (less than 4% of users have tweets on both sides). We apply the same principle to our compound score C. Then calculate our weighted compound per user in the same way sign(V)*C.

Visualization

Vocabulary score

We can see on the different maps in the appendix, the vocabulary used by the users. Furthermore, we can clearly see the variation between #ImpeachmentHearing and #ImpeachingHearing, the first one being more neutral and the second pro-impeachment. On the #ImpeachmentHearing vocabulary heatmap, using only our vocabulary score, we clearly recognize the US election map from the 2016 election. Thus, people uses the vocabulary according to the state political orientation. [Heatmaps]

For #ImpeachmentHearing:

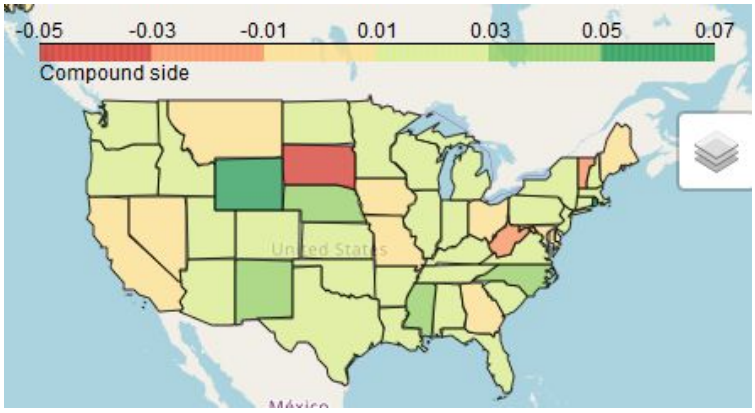


Weighted compound score

The orientation of a state about impeachment is not simply the political color. The overall Twitter users that tweeted about the impeachment hearing are for impeachment.

Further analysis on the validity of this map is tricky as we do not have any validation data as we did not find trustworthy polls per state on the impeachment. [heatmaps]

For #ImpeachmentHearing:



Limitations

Twitter is not only about text, many users only posts photos/gif/memes to convey an opinion. Thus our analysis based on text can not grasp the meaning of these tweets. Only a fraction of people from each state tweeted about this topic. [Percent of population tweeting about this #]

The results for WA are not assured as there might be a high classification error rate for this state.

Further analysis:

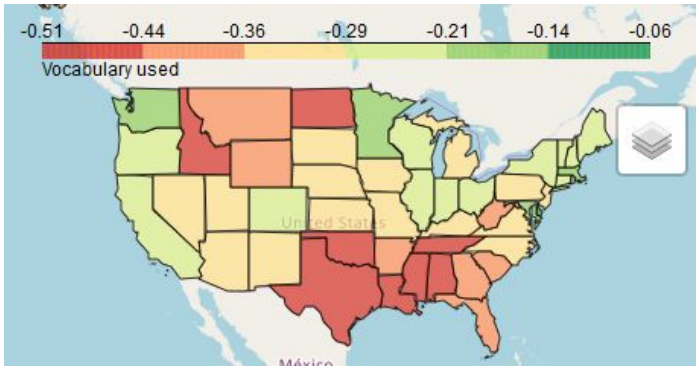
Our pipeline is fully operational to run a new analysis on similar events, from the api request to the heatmap. The only parameter we may want to adjust is the list of pro/anti words as the vocabulary and names involved in these events might differ from the hearing of mid-november.

The United States House of Representatives are elected by electoral districts. Thus we would need a deeper analysis on the location and try to look for a city.

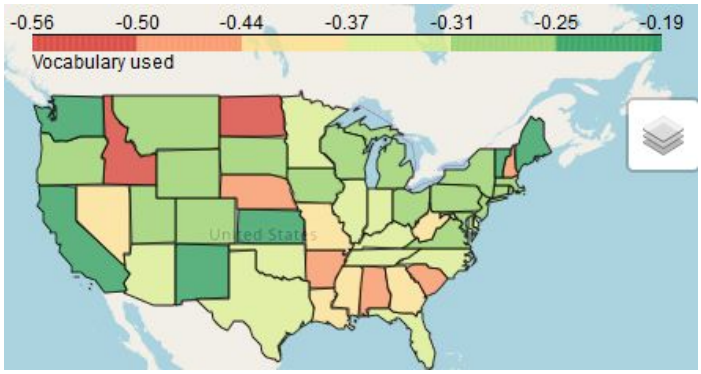
Appendix:

Heat map of the vocabulary coefficient: [to report]

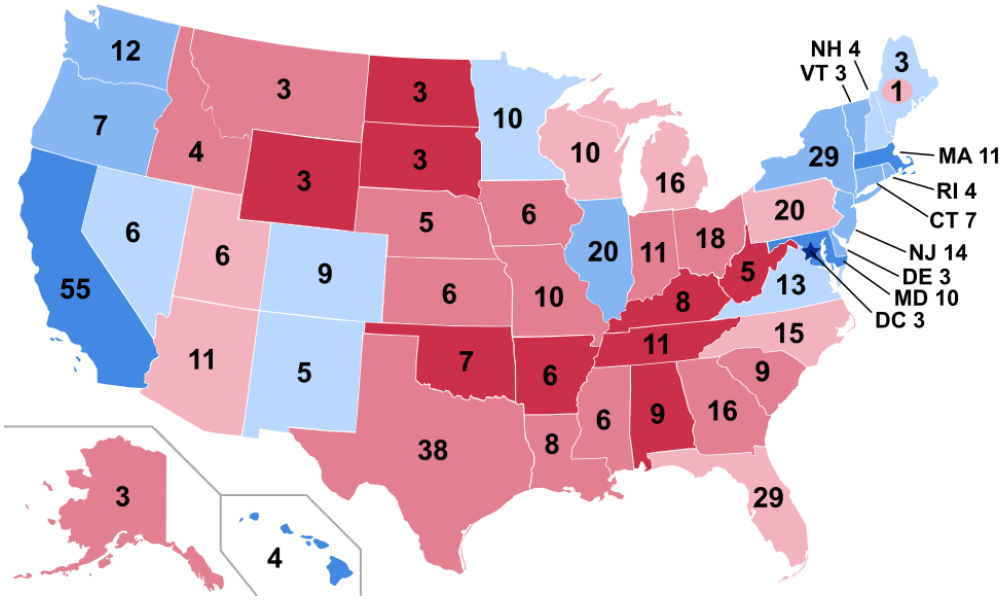
For #ImpeachmentHearing:



For #ImpeachingHearing:

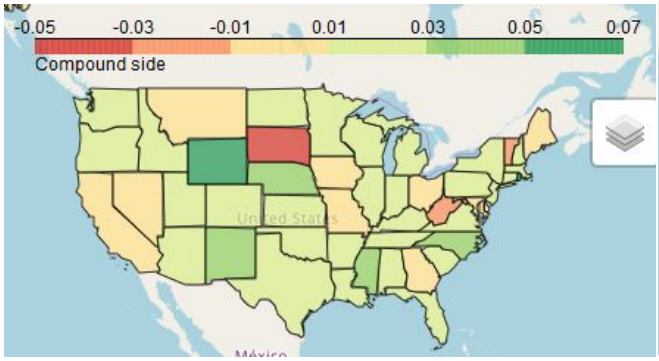


2016 election map:

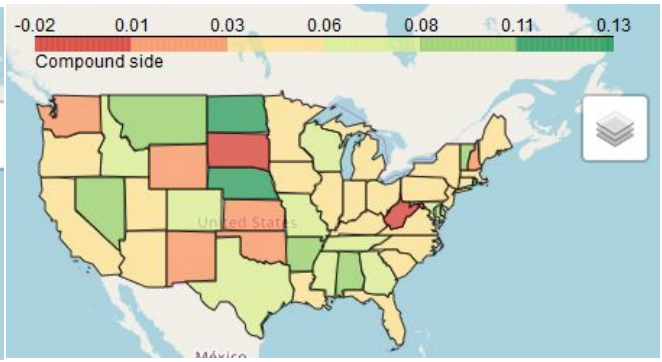


Heat map of the compounded weighted coefficient: [to report]

For #ImpeachmentHearing:



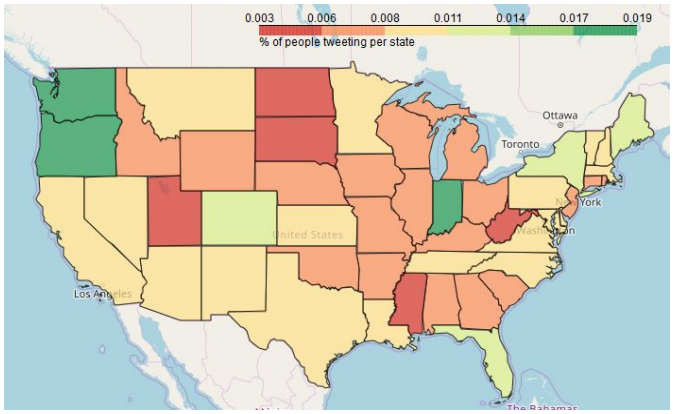
For #ImpeachingHearing:



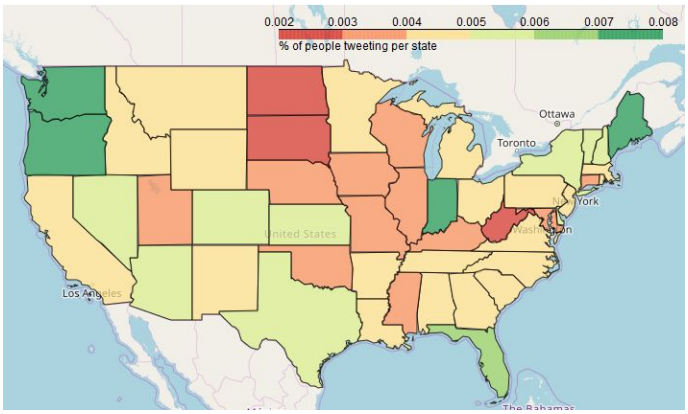
#ImpeachingHearing is effectively the most pro-impeachment #

Percent of population tweeting about this #: [to report]

For #ImpeachmentHearing (range 0.003 - 0.019%):



For #ImpeachingHearing (range 0.002 - 0.08%):



Population of the state : Estimation of 2019
https://simple.wikipedia.org/wiki/List_of_U.S._states_by_population

Keep in mind the classification error is high for WA, IN and maybe OR. (IN and OR as 'in' and 'or' are words and might have been used in the location description.)