Political Idiolects

Arjun Chadha

Introduction

- Highlight patterns and differences between Republican and Democratic ways of expression
- Can we train a model to accurately predict a political lean?
- Twitter as the SMS of the internet

Selecting Officials

Dealing with Bias - choosing Elected Officials

Criteria

Active on Twitter and was active during this past election

1

Fair representation of their party

2

Approved Officials

Republican

@GOPLeader / Kevin McCarthy
@SteveScalise / Steven Scalise
@Jim_Jordan / Jim Jordan
@RandPaul / Rand Paul
@GOP / Official Republican twitter
@Mike_Pence / Mike Pence
@MarshaBlackburn / Marsha Blackburn
@tedcruz / Ted Cruz
@marcorubio / Marco Rubio
@DanCrenshawTX / Dan Crenshaw
@LindseyGrahamSC / Lindsey Graham

Democrat

@BernieSanders / Bernie Sanders
@JoeBiden / Joe Biden
@ewarren / Elizabeth Warren
@PeteButtigieg / Pete Buttigieg
@KamalaHarris / Kamala Harris
@JohnDelaney / John Delaney
@amyklobuchar / Amy Klobuchar
@MichaelBennet / Michael Bennet
@AOC / Alexandria Ocasio-Cortez
@SpeakerPelosi / Nancy Pelosi
@TheDemocrats / Official Democratic twitter

Omitted Officials

Politicians that were considered but omitted because they would potentially add noise or skew the model:

Mike Bloomberg

Omitted because:
Has his foot in both camps

Donald Trump

Omitted because:
General inconsistency and abnormal tweeting behaviours

Mitt Romney

Omitted because:
First Republican to break away from Trump; acted as a litmus test for other Republicans

Tulsi Gabbard

Omitted because:
Too much of a band-wagoner; still has yet to make her position solid

Cleaning Process

Failure with Tweepy, issues pulling tweets with new character limit

Failure to pull data in JSON

Successfully went back to Tweepy

Regex pulls out ‘retweet’

Utilize Tweet preprocessor to clean emojis and BeautifulSoup to clean HTML waste

Data Format

- Dummy variables created from each of the official Twitter handles
- 22,000 tweets total, updated 3 different times in 2 months
- 2 separate Data frames for analysis, one for bagofwords, one for TF-iDF
- Target Variable = Is_Republican (1 or 0)

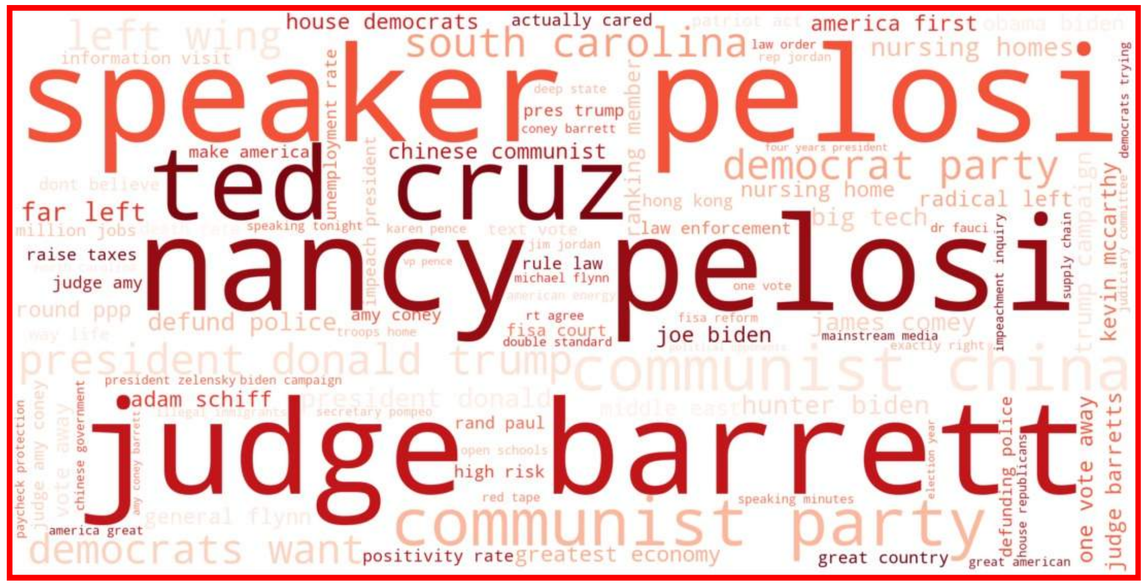
Preliminary Modeling

- Logistic regression worked best
- Tried different forms of vectorizing
- TF-iDF_nostem_stop used because it overfit the least and it still provided good test accuracy

	Training Accuracies	Test Accuracies
CountVecNoStemorStop	0.941939	0.817818
CountVecNoStem_Stop	0.934182	0.814182
CountVecNoStop_Stem	0.937091	0.792727
CountVecStop&Stem	0.930545	0.792364
TF-idfNoStemorStop	0.880121	0.802909
TF-idfNoStem_Stop	0.886545	0.811273
TF-idfNoStop_Stem	0.876485	0.794909
TF-idfStop&Stem	0.883394	0.792364

Results - One word tokens

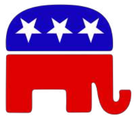
Results - Two word tokens



Results Analyzed

1

Republicans heavily into state rights and campaigning whereas Dem's talk about country as a whole.



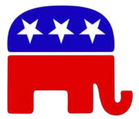
'florida'
'texas'
'pennsylvania'
'houston'
'carolina'

'everyone'
'build'
'plan'
'us'
'country'
'capitol'
'together'
‘soul of the nation’

Results Analyzed

2 Republicans are focused on the economic health and status of the country, whereas the Democrats are focusing on the people themselves and social issues.

'big tech'
'greatest economy'
‘small business’
‘unemployment rate’
‘great country’
‘american energy’
‘supply chain’
‘PPP’



'health care’
‘systemic racism’
‘young people’
‘working class’
‘black women’
‘working families’
‘community’
‘people’

Results Analyzed

3 **7 highly predictive tokens expressed by Democrats about voting, and none from the Republicans.**

'vote early' *'voting plan'* *'register vote'*

'early voting' *'voting mail'* *'voter suppression'*

Results Analyzed

4a Interesting to note that the Republicans talk a lot about the Democrats...

'dems' 'left' 'democrats' 'liberal' 'left wing' 'democratic party' 'democrats trying'

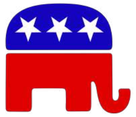
4b ...Not one democratic token is indicating towards the Republicans.

Results Analyzed

‘trump’
‘donald’
‘donald trump’

5

Differentiation in addressing the President



‘president donald’
‘president donald trump’
‘pres trump’

Results Analyzed

6 **Republicans seem to worry about the extremist faction of the opposing party a lot more than the Democrats.**

‘radical left’ ***‘far left’***

Results Analyzed

7 Republicans in general have been and were very vocal about China at the Republican National Convention so the word "chinese" and "china/china(')s" were used a lot and therefore propagated on twitter. “CCP” and “Communist China” are predictive words which are points of contention that Republicans have about the Chinese.

'RNC' *‘china/china(‘)s’* *‘chinese’*

‘CCP’ *‘communist party’* *‘communist china’*

Results Analyzed

8 **Democrats use lots of action words to highlight and critique the presidency as a whole.**

'need' *'failed'* *'crisis'* *'everyone'*

'trump' *'future'* *'change'* *'better'* *'must'*

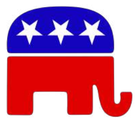
'trump failed' *'deserve president'*

Results Analyzed

9

Indication of values/patterns that make sense.

'god'
'government'
‘calling’
‘people’
‘praying’
‘law enforcement’
‘due process’
‘american energy’
‘deep state’
‘red tape’
‘law and order’
‘way of life’



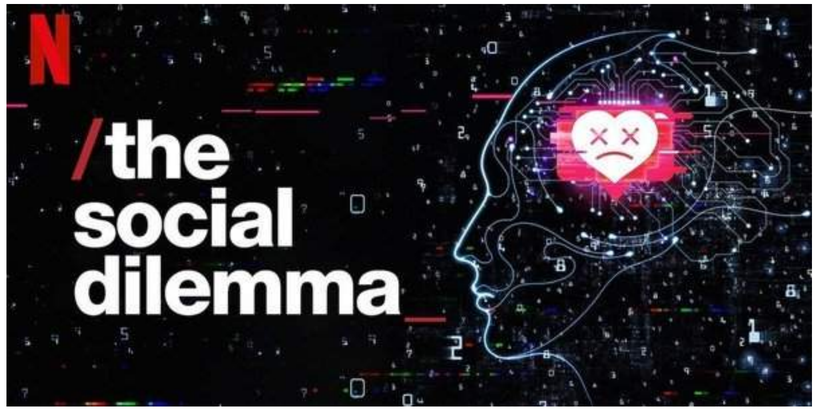
'liberty'
'climate crisis’
'plan'
'racial justice’
'reproductive rights'
'leadership'
'wealth'
'promise'
‘LGBTQ’
‘voices heard’
‘white supremacists’
‘gun violence’

Testing the political lean of other Members of Congress

Classification of 500 most recent tweets

	Democratic Tweets	Republican Tweets
Angus King	71.3%	28.7%
Donald Trump	17.4%	82.6%
Mitt Romney	32.3%	67.8%
Barack Obama	77.6%	22.4%

Next Steps



- Potential bot detection on social media platforms
- Enhance models with BERT using transfer learning

Thank You!

Arjun Chadha

arjun.rj.chadha@gmail.com

linkedin.com/in/arjun-chadha-ds