US ELECTION 2020

Tweets

Team:
Yuliia Nikolaenko
Mihaela Grigore
Lara Wagner (Wrangler)

DATASET and Features

TRUMP Dataset
(tweets containing #trump)

970919 tweets

3 weeks

BIDEN Dataset
(tweets containing #biden)

776886 tweets

21 Features

<u>Related to tweets:</u>
ID, Tweet, Likes, Retweet, Source ...

<u>Related to users:</u>
ID, Name, Description, Location, Join date, country, state ...

Date Today

~~21 01 2021~~

November 2020						
M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Monday 9th Nov 2020

Campaign management Team for Biden

Strategy Pointers for further campaigning

We are in the campaigning period and gathered the previous **3 weeks of data** from Twitter.

We want to get insight from this data to help **increase public support** (which we take as a proxy for how they will vote) **for Biden**.

Strategy Pointers for further campaigning

1. How is the discourse around the elections and candidates formed?
2. Which actors/events/statements are represented in the tweets?
3. In which states is Biden behind Trump?
   (based on sentiment scores of tweets)
4. What is the most discussed topic in each state and how do people feel about it?

TABLE OF CONTENTS

★ 01 ★
CLEANING AND PREPROCESSING

★ 02 ★
EDA

★ 03 ★
SENTIMENT ANALYSIS

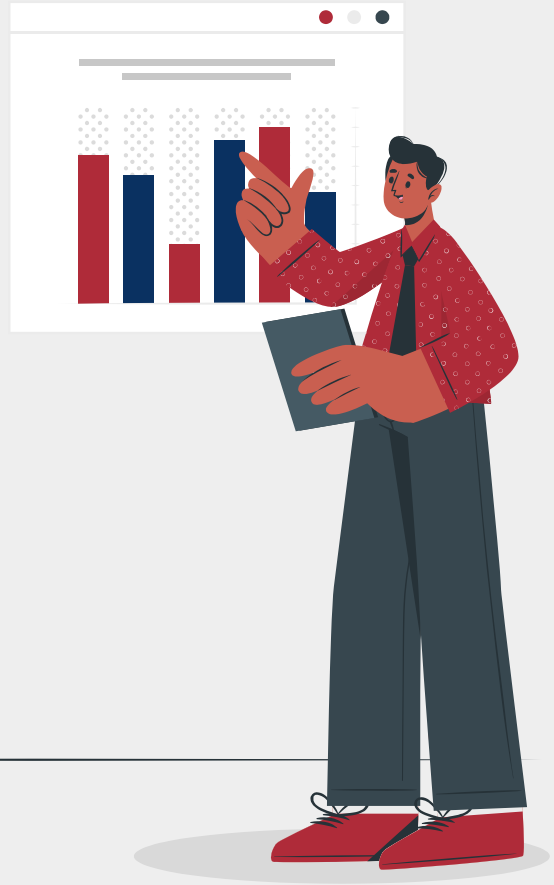
★ 04 ★
TOPIC MODELLING AND PREDICTION

★ 05 ★
TOPICS DISTRIBUTION OVER STATES

★ 01 ★

CLEANING AND PREPROCESSING

Preparing data for the analysis



FINAL DATASET

Users based in USA

Remove duplicates
Tweets containing both #trump and #biden

TRUMP

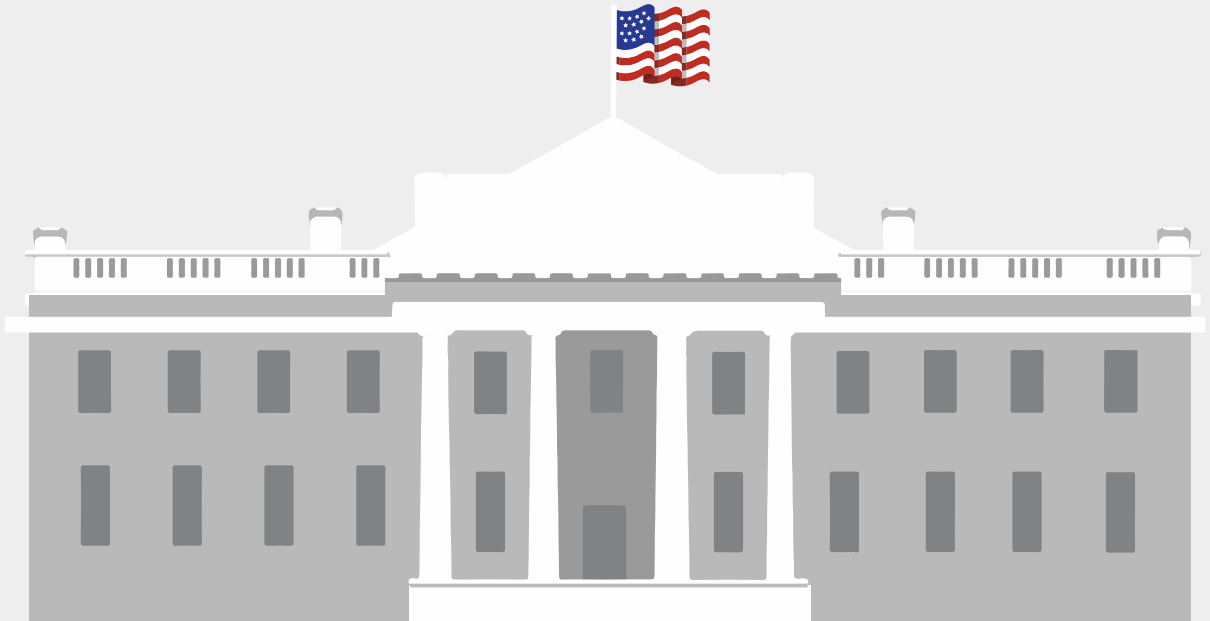
81884 tweets

BIDEN

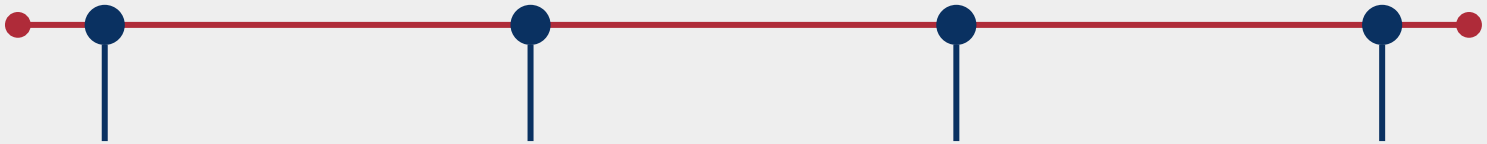
70492 tweets

Final dataset: 152376 tweets

US tweets for both candidates



PREPARING THE TEXT FOR THE ANALYSIS



LOWERCASING

REMOVING LINKS, PUNCTUATION, STOPWORDS AND REGEX

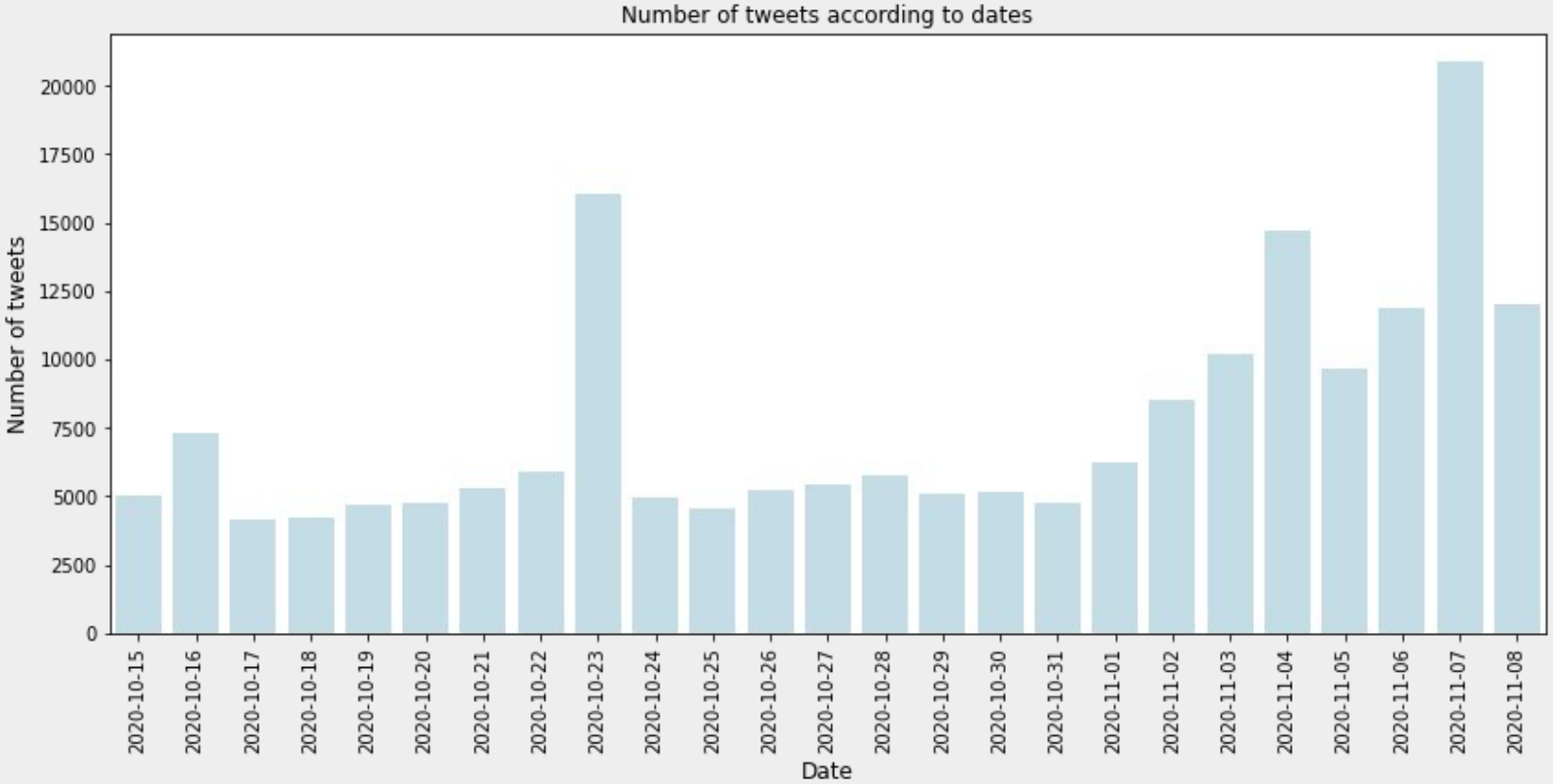
LEMMATIZING

TOKENIZATION

02

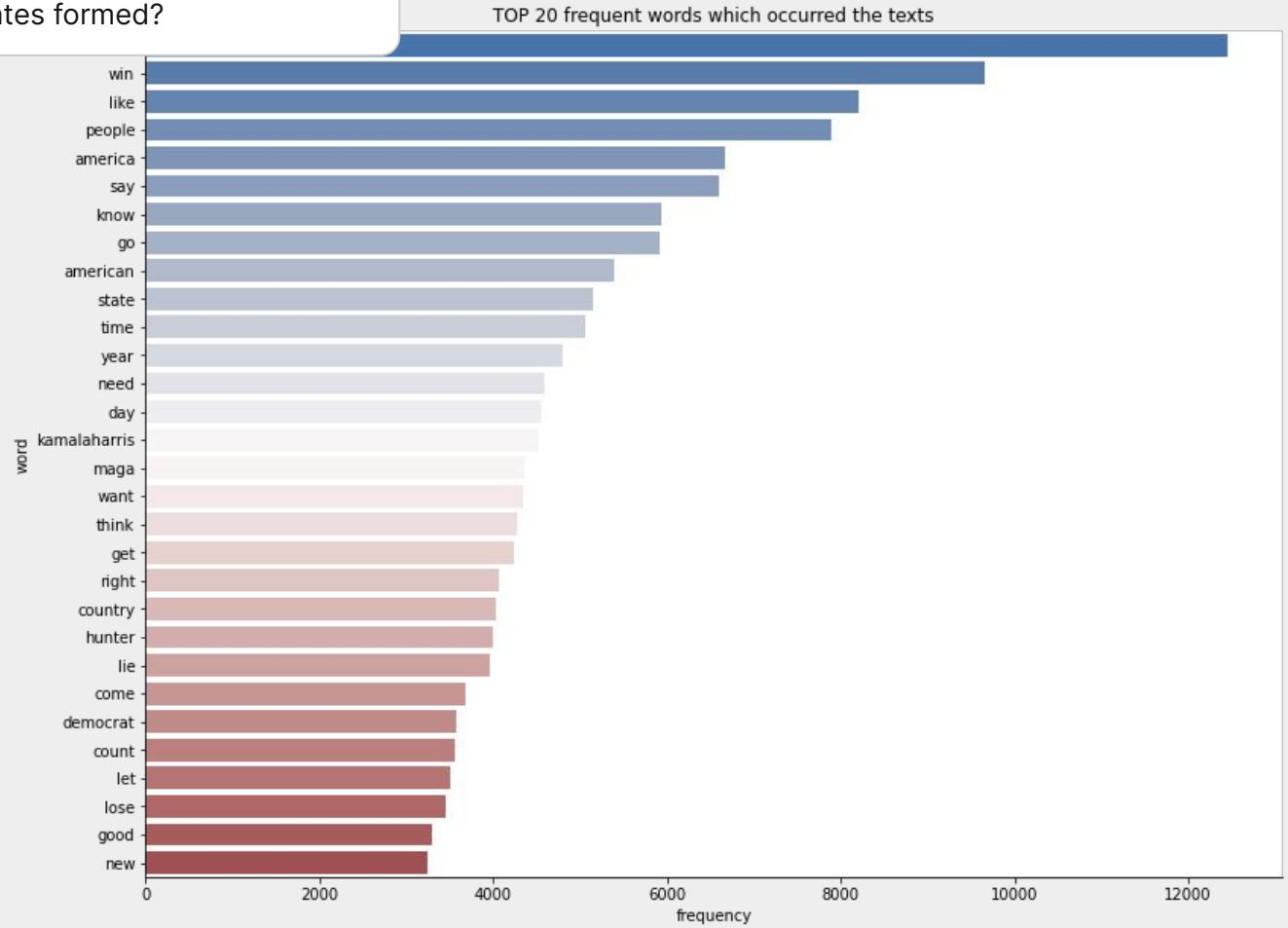
EDA

Exploring the discourse around the elections

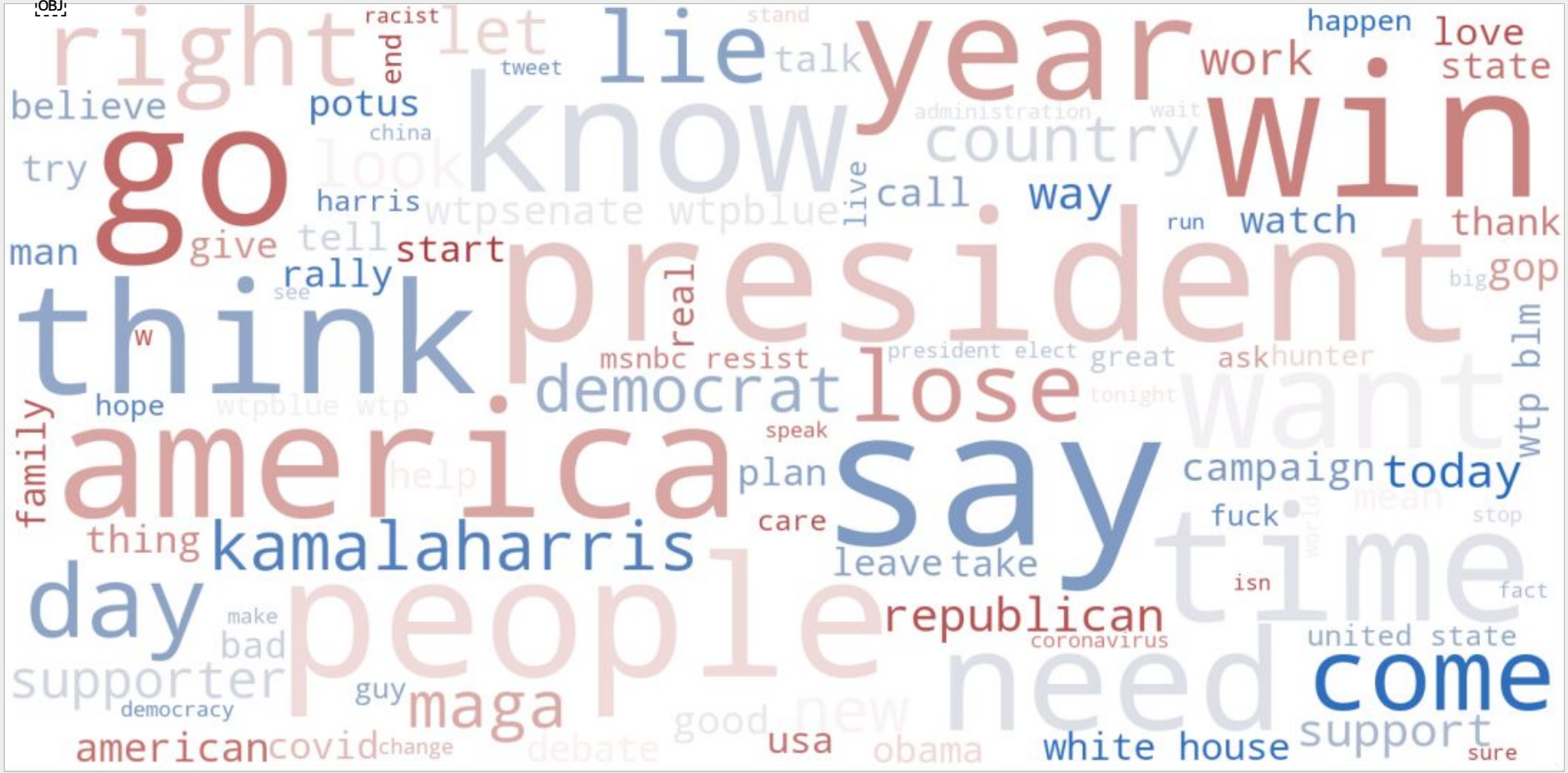

Number of tweets according to dates
Number of tweets
0
2500
5000
7500
10000
12500
15000
17500
20000
2020-10-15
2020-10-16
2020-10-17
2020-10-18
2020-10-19
2020-10-20
2020-10-21
2020-10-22
2020-10-23
2020-10-24
2020-10-25
2020-10-26
2020-10-27
2020-10-28
2020-10-29
2020-10-30
2020-10-31
2020-11-01
2020-11-02
2020-11-03
2020-11-04
2020-11-05
2020-11-06
2020-11-07
2020-11-08
Date

1

How is the discourse around the elections and candidates formed?



Wordcloud of TOP 100 frequent words for all tweets

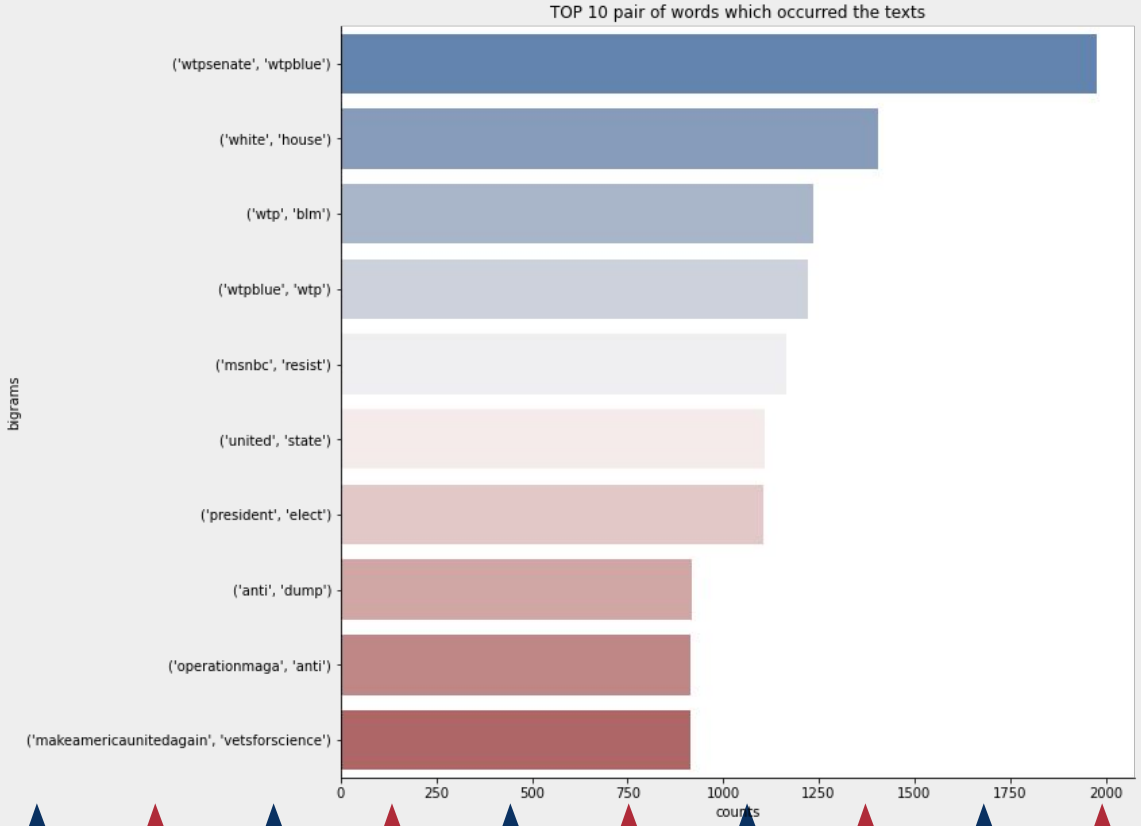


Wordcloud of TOP frequent words for Trump tweets

Wordcloud of TOP frequent words for Biden tweets

2

Which actors/events/statements are represented in the tweets?



★ 03 ★

SENTIMENT ANALYSIS

Popular methods for sentiment analysis

	TextBlob	VADER	Flair
Overview		VADER (Valence Aware Dictionary for Sentiment Reasoning), 2014	character-level **LSTM** neural network
Relevant Output	Something ∈ [-1, 1]		
Advantages (+) Disadvantages (-)	Bag of words + Subjectivity (factual - opinionated) - does not have heuristics of VADER	+ Bag of words | empirically validated by multiple human judges | for **microblog-like** contexts | heuristics (negations & intensifiers) - typos <=> OOV	+ recognizes typos (no OOV), negations ('not like it'), intensifiers ('so easy') - training is resource intensive. Pre-trained model on IMDB & is extremely polarizing
Interpretation	neg (<0.0), pos (>0.0), neg(rest)		
References	official website	github	github

Whom to trust

We ran all 3 and looked at how much they agree / disagree on the sentiment assigned per tweet

Agreement on sentiment

Consensus TextBlob - VADER	56%
Consensus TextBlob - Flair	30%
Consensus VADER - Flair	36%

Analysis on labels (neg, pos, neutral)

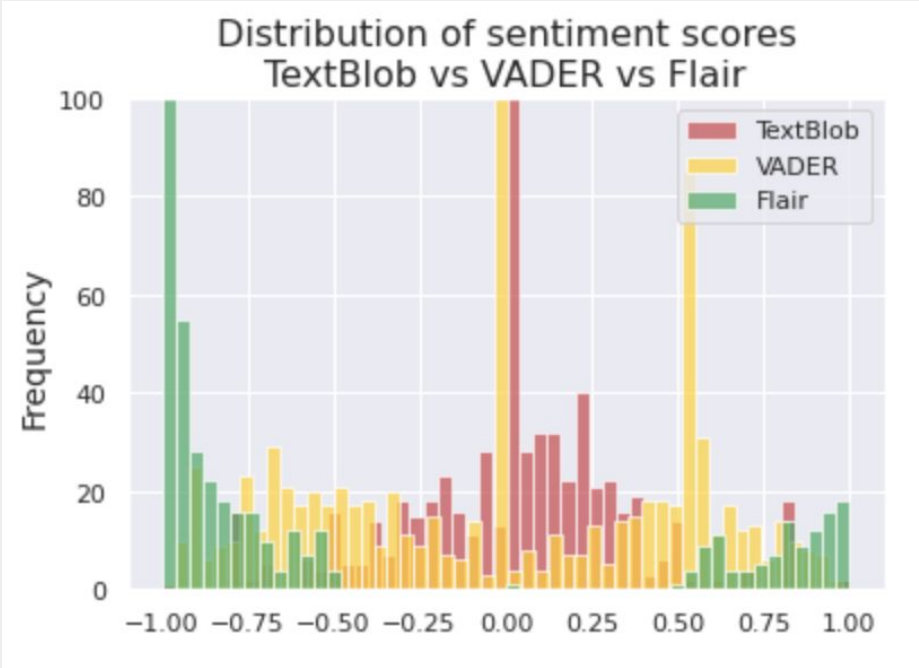
tweet	blob	vader	flair
#NBCboycott @NBCNews fire #ChuckTodd and move the town hall with #trump	neutral	negative	positive
All these people upset because #IceCube is working with #trump! Maybe the problem is you see this rapper as your hero. Time to upgrade your hero's to #MLK #nelsonmandela @Kaepernick7	neutral	negative	positive
Head Thread. #Trump & co. keep committing all manner of Fraud, Libel, Corruption, etc, to keep afloat. #HunterBiden #ButHisEmails #RudyColludy #Burisma https://t.co/QpWaukVErn	neutral	negative	positive
@business @bsurveillance : @tomkeene & @flacqua : #Stephanie you cant compare #Trump's performance in 2016 to now. Why? He had no record in public office then. He does now!!	neutral	negative	positive
@FLOTUS @realDonaldTrump Smh...no doubt, #donaldtrump is the trick.\n#HushMoney\n#Individual1\nhttps://t.co/jVJt2O5oqJ	neutral	negative	positive
@FLOTUS @realDonaldTrump This is how #donaldtrump gets money in his pockets...steals from kids with cancer. \n#TrumpIsACriminal\n#TrumpForPrison\n#TrumpIsAPos\nhttps://t.co/YbMVaOgcY2	neutral	negative	positive
@realDonaldTrump No doubt...garbage.\nDesperation from #donaldtrump & his lackeys.\n#TrumpIsANationalDisgrace \n#TrumpIsALaughingStock \n#TrumpLiesEverytimeHeSpeaks \n#TrumpIsAPos\nhttps://t.co/Oqdv6Oqj8n	neutral	negative	positive
@GOP @realDonaldTrump This is how #donaldtrump gets money in his pockets...steals from kids with cancer. \n#TrumpIsACriminal\n#TrumpForPrison\n#TrumpIsAPos\nhttps://t.co/YbMVaOgcY2	neutral	negative	positive

Further exploration of sentiment score

Observations:

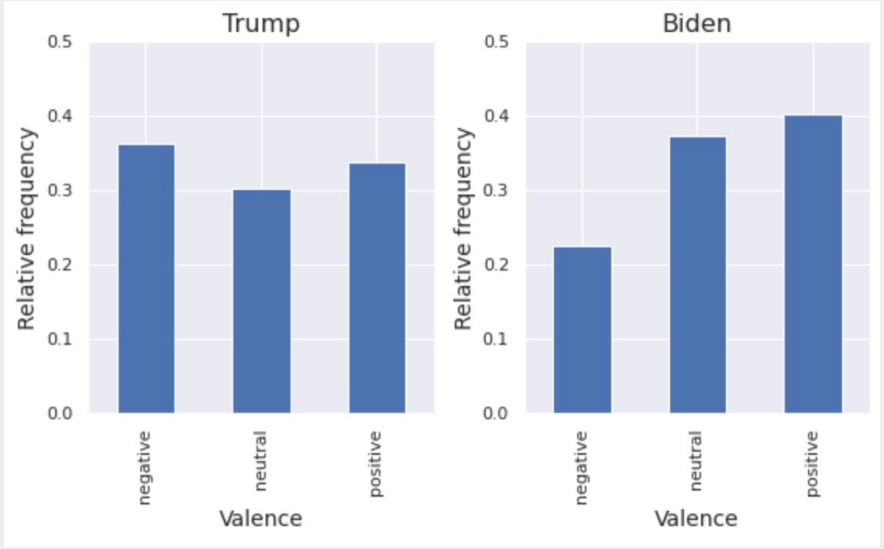
- Flair seems very polarizing (by comparison)
- TextBlob and VADER have a ‘strong’ neutral category
- TextBlob and VADER look plausible

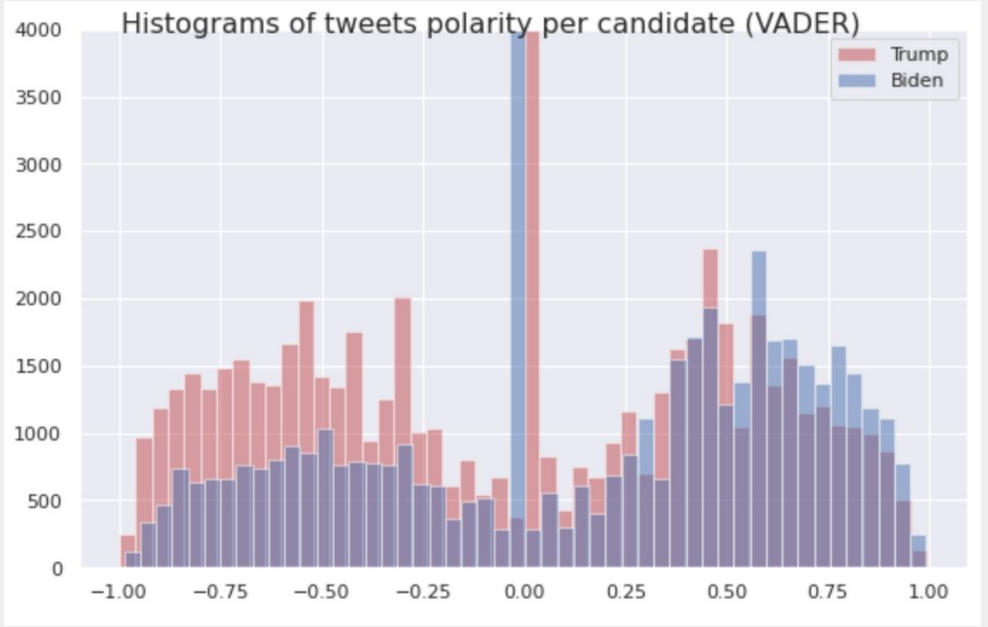
Decision: we chose VADER



Analysis on score ∈ [-1,1]

Overall sentiment in Biden vs Trump tweets





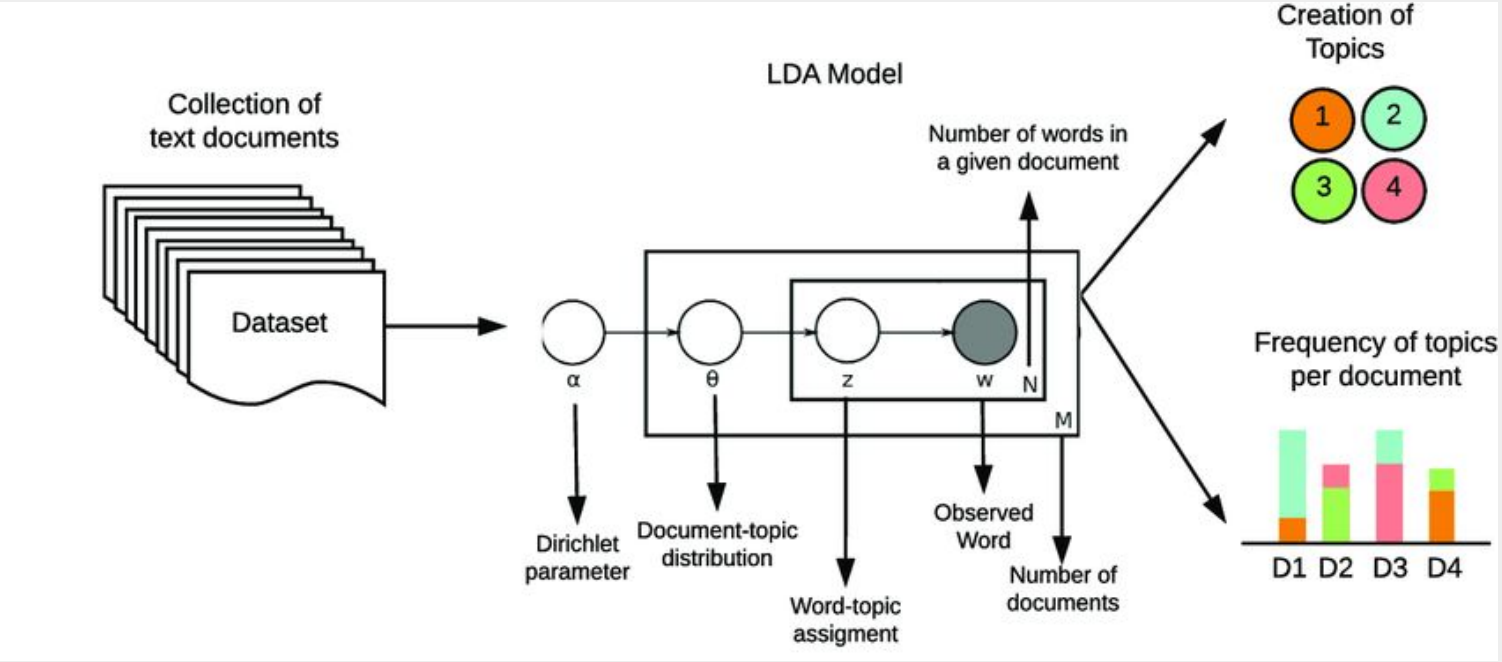
Sentiment label: categorical ∈ {negative, neutral, positive}

Sentiment score: continuous ∈ [-1,1]

★ 04 ★

TOPIC MODELLING AND PREDICTION

Latent Dirichlet allocation MODEL



STRATEGY

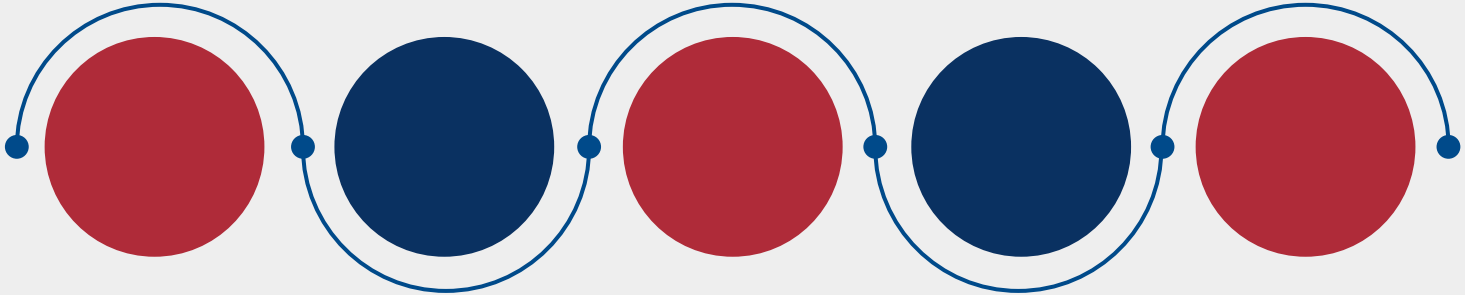
Step 1
Computing the coherence scores for the different number of topics

Step 2
Creating the model with optimal number of topics

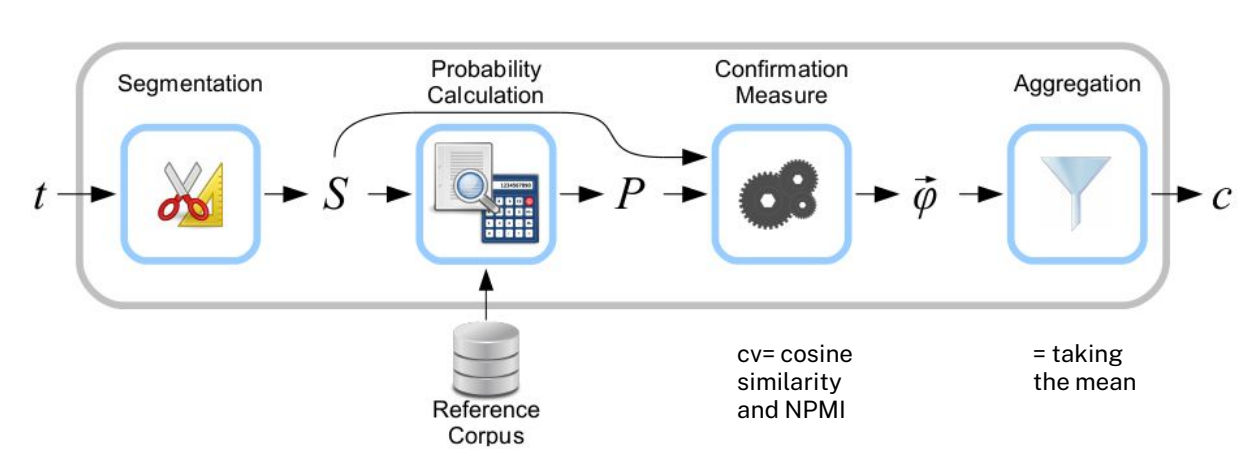
Step 3
Topics extraction and visualization

Step 4
Labeling the tweets related to the topics

Step 5
Predicting the topics for the new tweets



COHERENCE SCORES



https://rare-technologies.com/what-is-topic-coherence/

t : Topics coming in

S : Segmented topics

P : Calculated probabilities

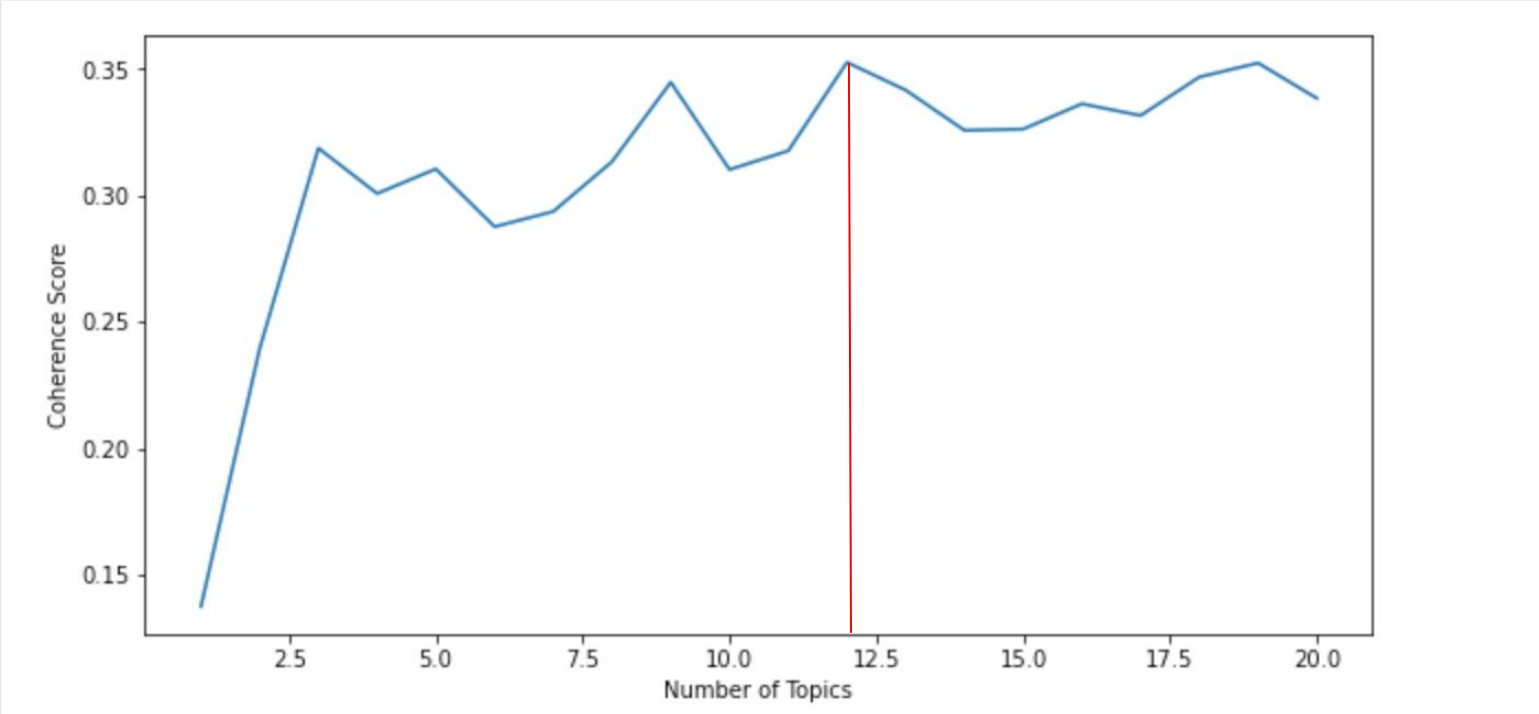
Phi vector : A vector of the “confirmed measures” coming out from the confirmation module

c : The final coherence value

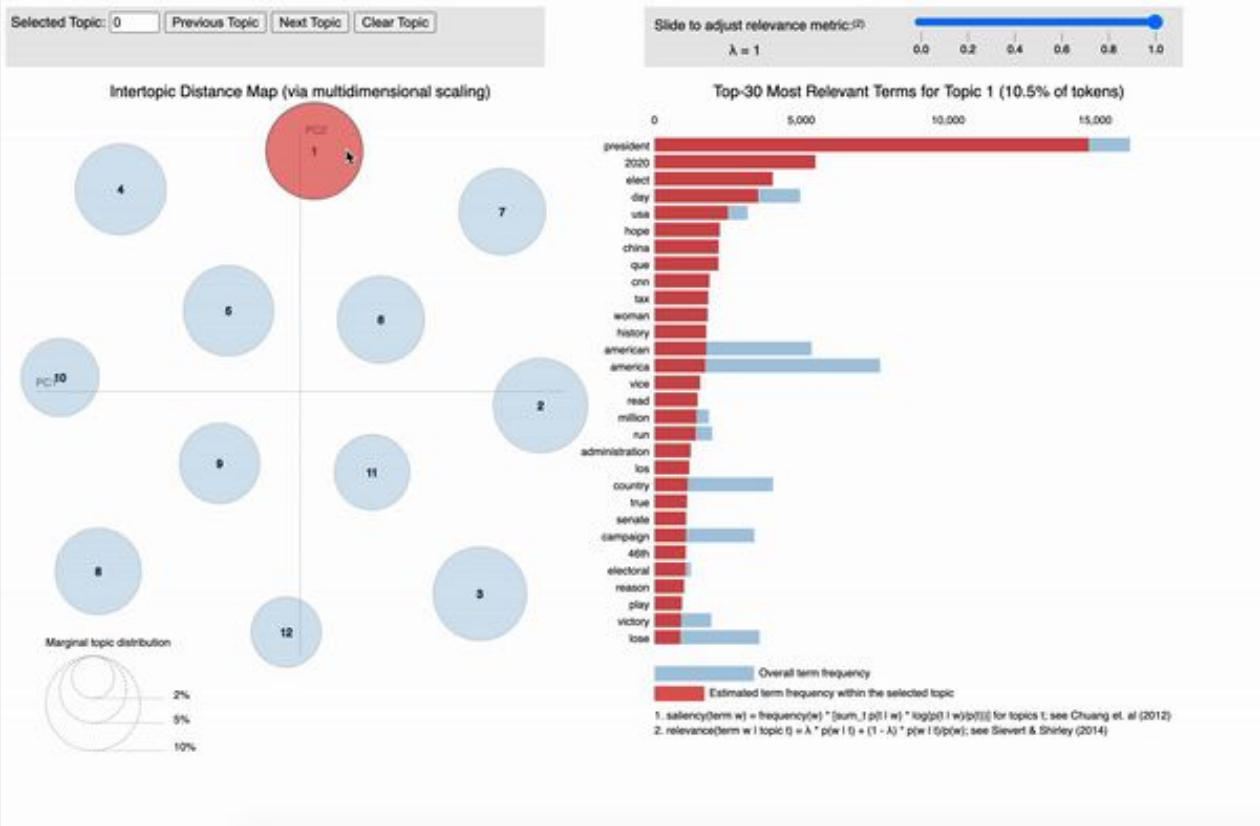
(Something ∈ [-1, 1])

COHERENCE SCORES

How semantically coherent a model is



MODEL VISUALIZATION



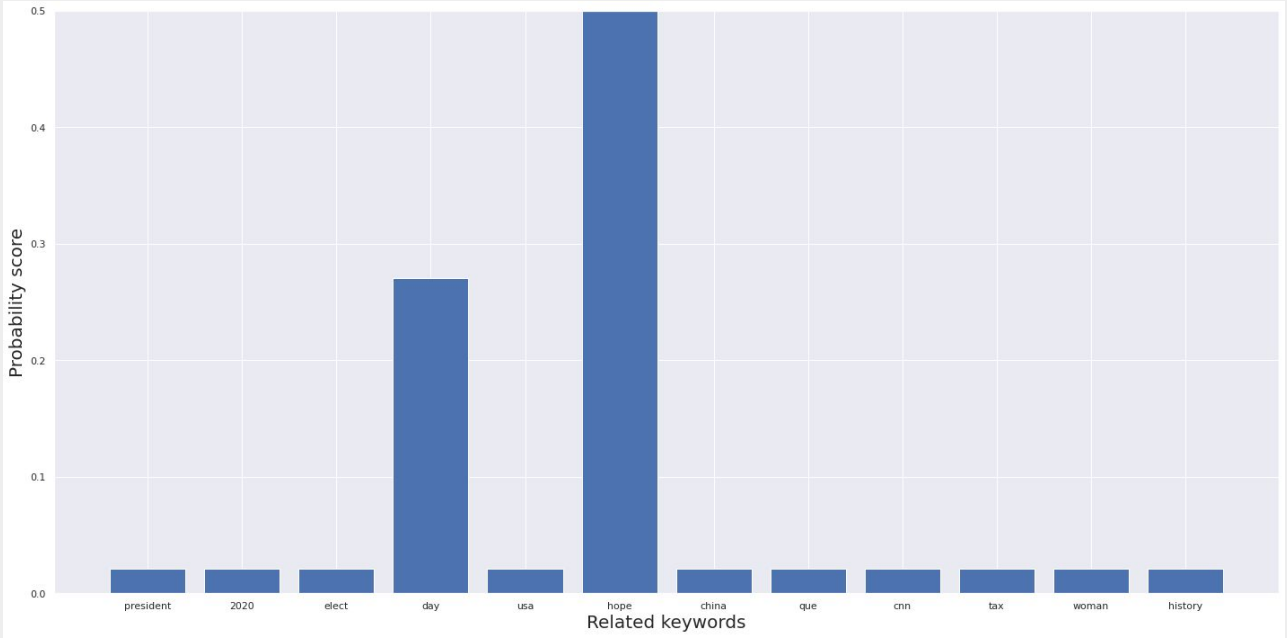
Try the interactive visualization on this link

Topic	Key words	Name assigned
1	'leader', 'corruption', 'nevada', 'gop', 'politics', 'video', 'real', 'georgia', 'ballot', 'hunter'	Election fraud
2	'united', 'medium', 'michigan', 'believe', 'look', 'people', 'country', 'news', 'need', 'state'	News
3	'people', 'mail', 'today', 'man', 'work', 'obama', 'american', 'count', 'pennsylvania', 'let'	Voting in Pennsylvania
4	'time', 'gon', 'matter', 'turn', 'thing', 'hear', 'kamala', 'win', 'life', 'good'	Biden and Kamala
5	'like', 'lose', 'poll', 'time', 'talk', 'debate', 'try', 'maga', 'know', 'win'	Trump losing
6	'tax', 'cnn', 'que', 'china', 'hope', 'usa', 'day', 'elect', '2020', 'president'	CNN
7	'arizona', 'party', 'florida', 'stop', 'covid', 'covid19', 'win', 'republican', 'democrat'	Covid
8	'racist', 'fraud', 'guy', 'say', 'speech', 'presidential', 'house', 'white', 'lie', 'presidentelect'	Trump speech
9	'case', 'claim', 'anti', 'feel', 'fuck', 'lead', 'blm', 'dump', 'new', 'come'	BLM
10	'red', 'lol', 'political', 'truth', 'child', 'person', 'demvoice1', 'bluewave', 'care', 'end'	Biden political program
11	'msnbc', 'black', 'thank', 'support', 'wtpsenate', 'year', 'love', 'harris', 'america', 'kamalaharris'	Biden supporting
12	'campaign', 'democracy', 'watch', 'debates2020', 'family', 'supporter', 'great', 'wtpblue', 'like', 'think'	Biden supporting

PREDICTING TOPICS FOR THE NEW TWEETS

NEW TWEET: "Biden is the best president for this country vote for him"

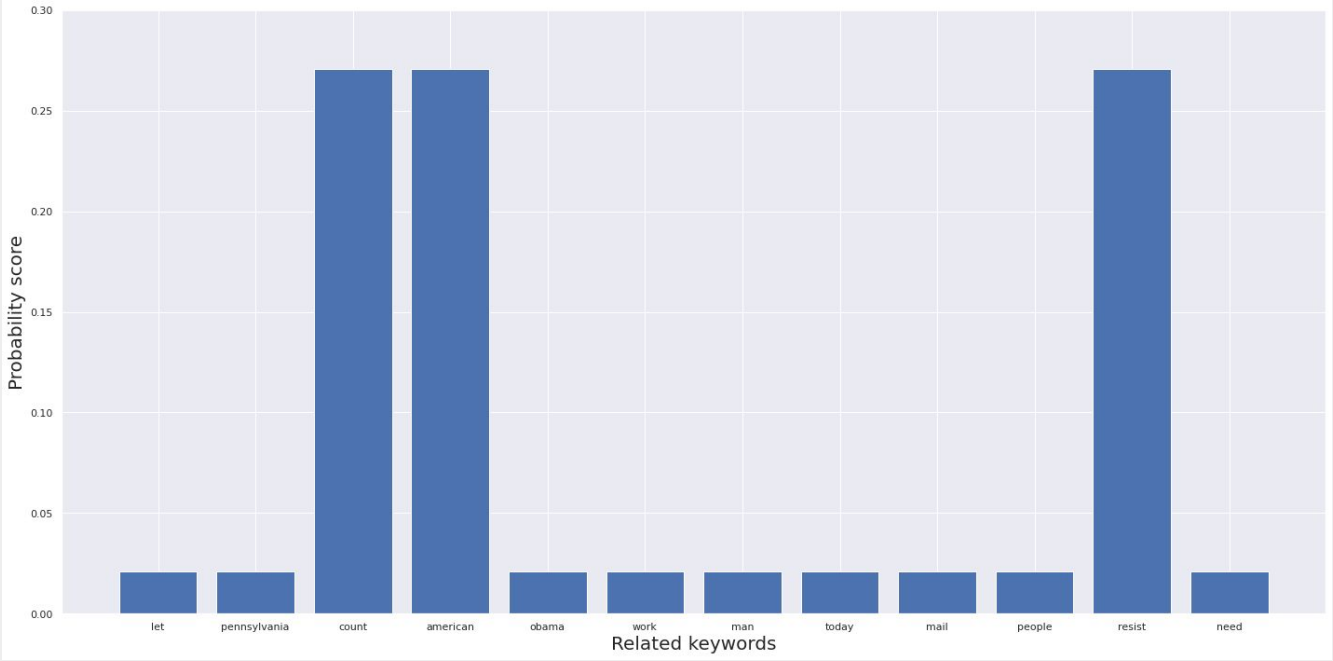
TOPIC 6



PREDICTING TOPICS FOR THE NEW TWEETS

NEW TWEET: "Trump is the worse thing happened to USA"

TOPIC 2



★ 05 ★

TOPIC DISTRIBUTION OVER STATES

3

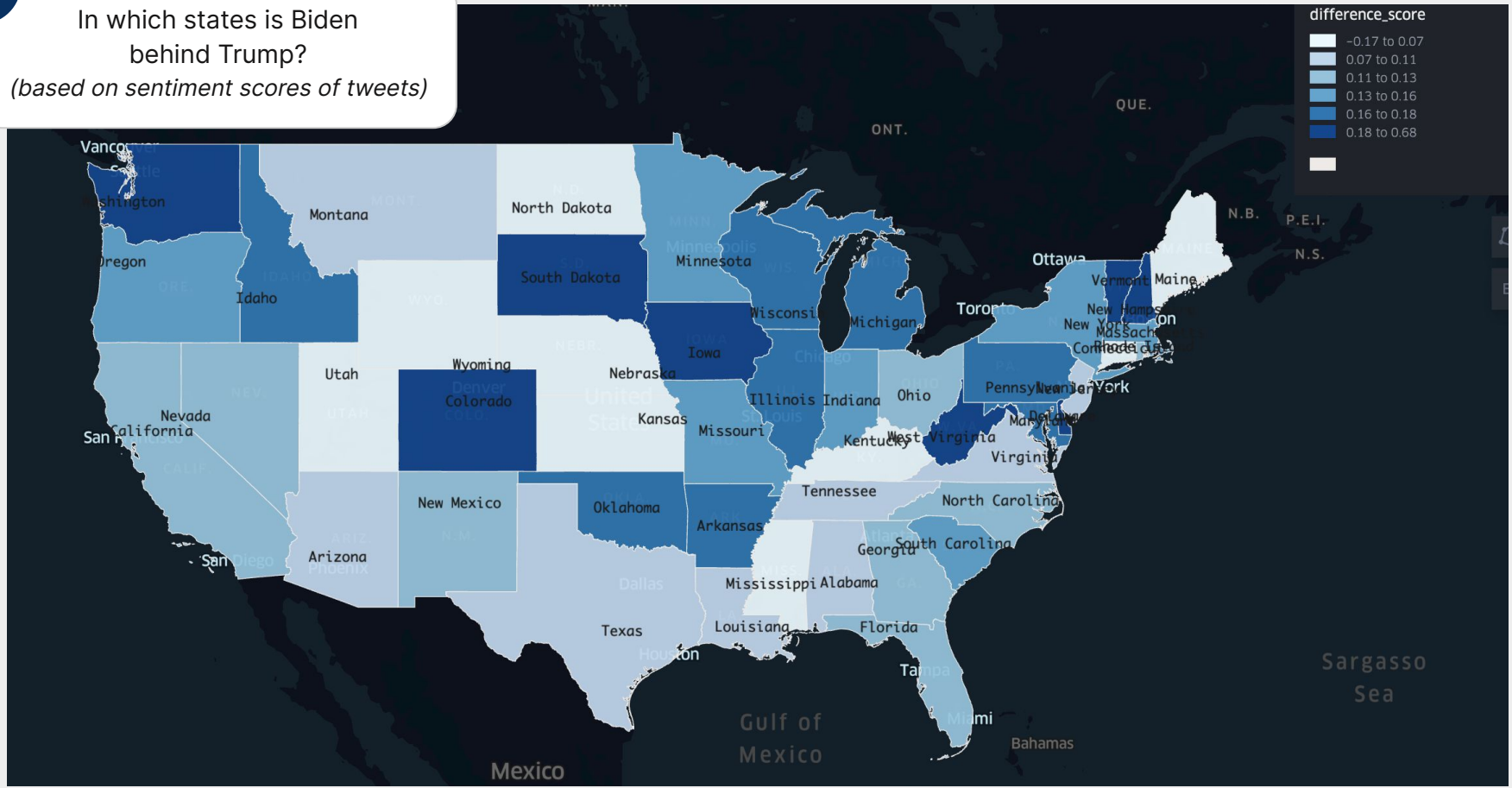
In which states is Biden behind Trump?
(based on sentiment scores of tweets)

Country	President	Sentiment Score	Difference
Mississippi	Trump	0.046	-0.097
	Biden	-0.050	
Utah	...	...	...

3

In which states is Biden behind Trump?

(based on sentiment scores of tweets)

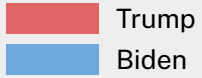


Explore interactive map

3

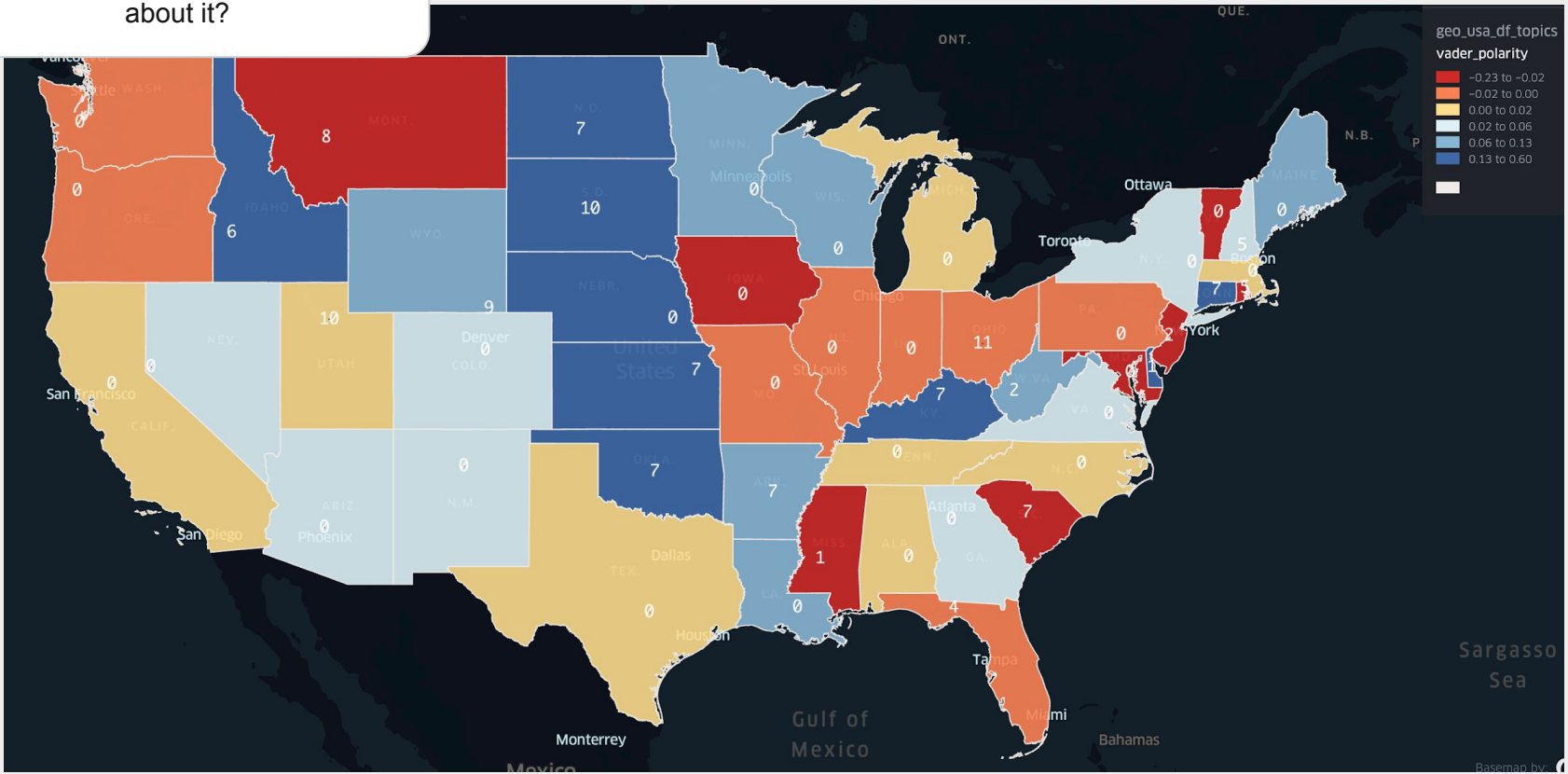
In which states is Biden behind Trump?
(based on sentiment scores of tweets)

States	Leading Party
Maine	-0.169
Wyoming	-0.120
Mississippi	-0.097
Utah	-0.046
North Dakota	-0.001
Kansas	0.014



4

What is the most discussed topic in each state and how do people feel about it?



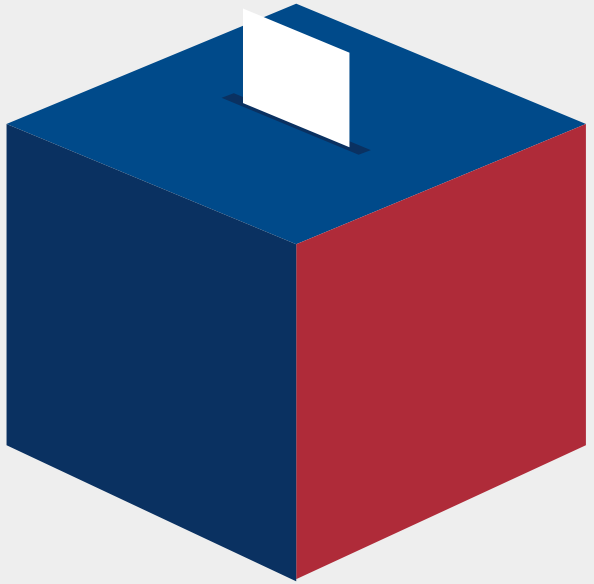
Explore interactive map

4

What is the most discussed topic in the intervention states and how do people feel about it?

States	Nr	Topic	Polarity	Sentiment Score
Maine	1	Election Fraud	Neutral	0.073
Wyoming	10	Biden political program	Positive	0.121
Mississippi	2	News	Negative	-0.043
Utah	11	Biden supporting	Neutral	0.015
North Dakota	8	Trump Speech	Positive	0.204
Kansas	8	Trump Speech	Positive	0.202

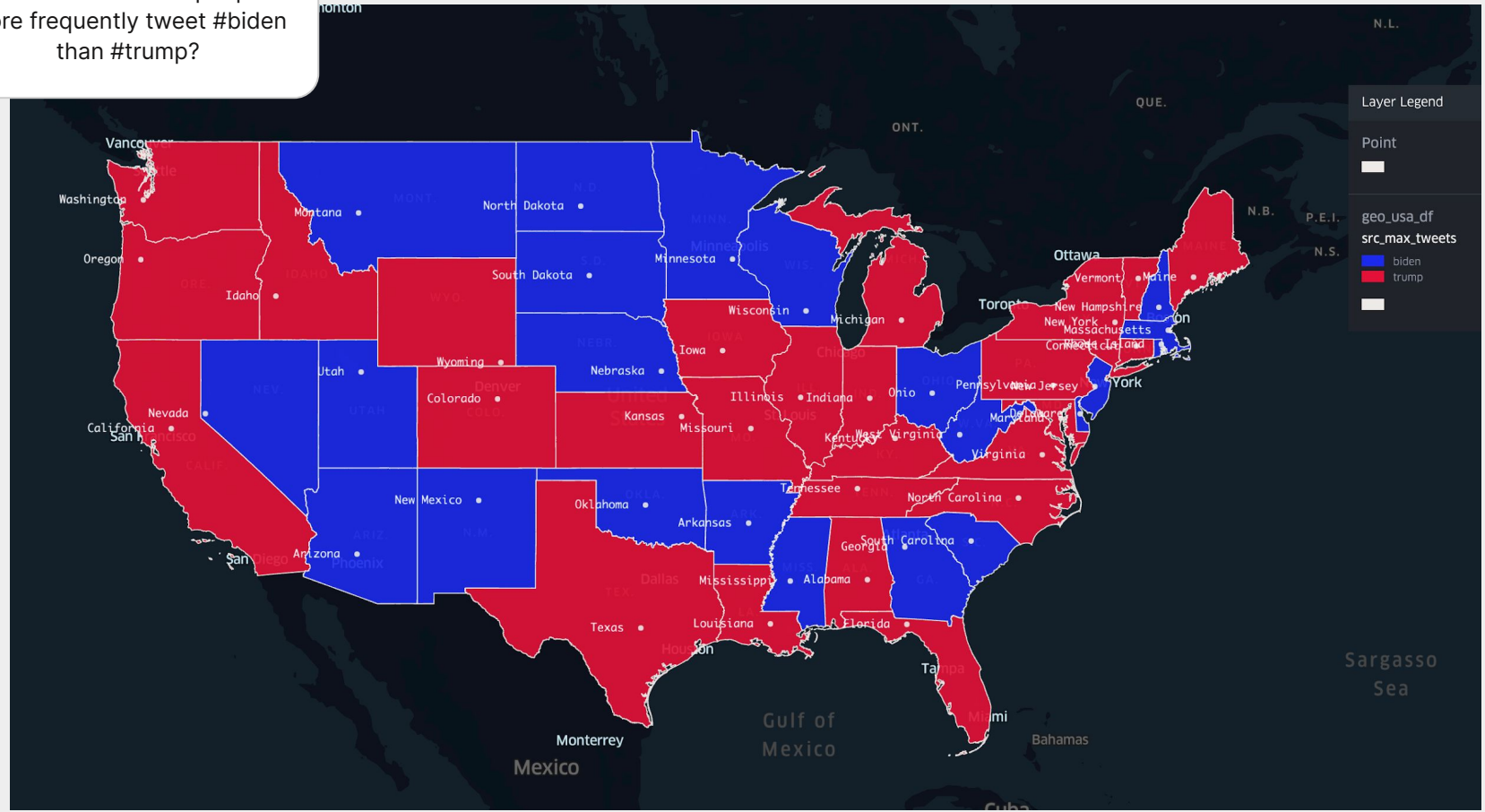
THANKS FOR YOUR ATTENTION!



1

In which states did people more frequently tweet #biden than #trump?

Distribution of tweets mentioning trump vs. biden



2

In which states is Biden ahead of Trump?

States	Leading Party
Maine	53.1%
Wyoming	70.4%
Mississippi	57.6%
Utah	58.2%
North Dakota	65.5%

States	Leading Party
Delaware	58.8%
West Virginia	68.6%
South Dakota	61.8%
Vermont	66.4%
Colorado	55.4%

3

In which states is Biden behind Trump?
(based on sentiment scores of tweets)

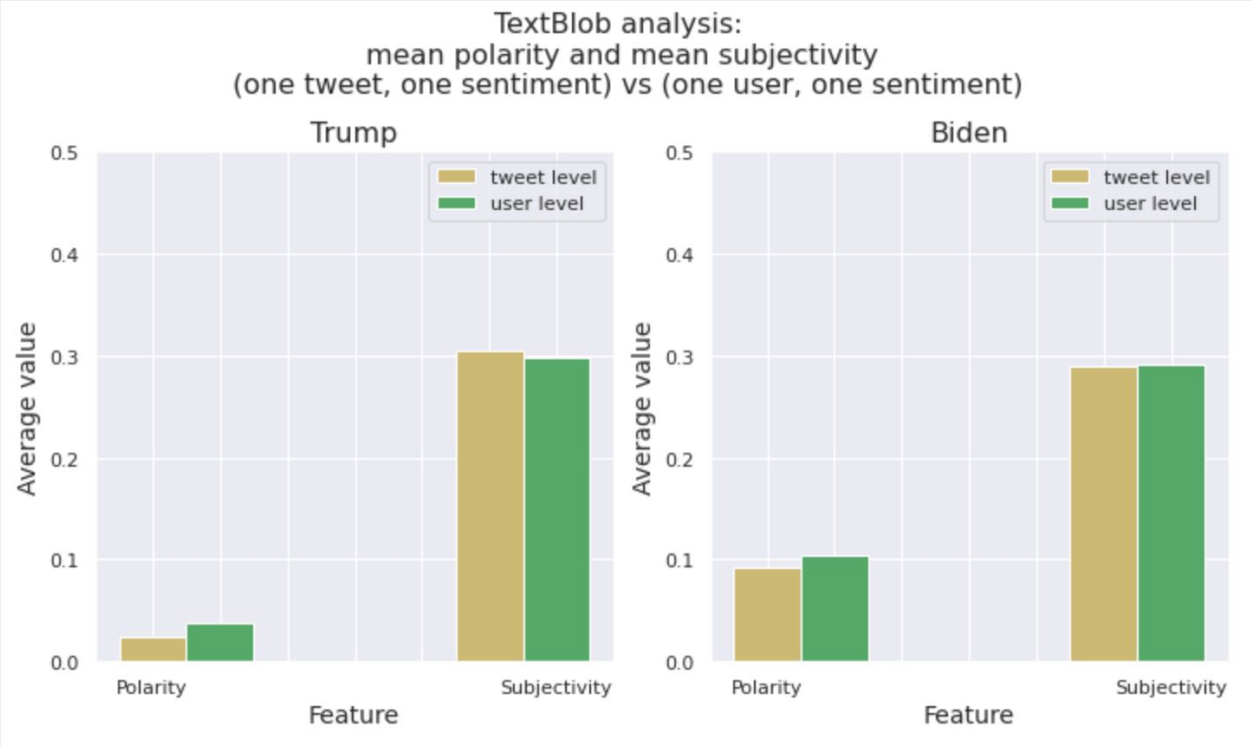
States	Leading Party
Maine	-0.169
Wyoming	-0.120
Mississippi	-0.097
Utah	-0.046
North Dakota	-0.001
Kansas	0.014

States	Leading Party
Delaware	0.679
West Virginia	0.326
South Dakota	0.305
Vermont	0.222
Colorado	0.216
Alaska	0.211

3

What topic is each state more interested in?

States	Nr	Topic	Sentiment Score	Won
Maine	1	Election Fraud	0.073	Biden (53.1%)
Wyoming	10	Biden political program	0.121	Biden (70.4%)
Mississippi	2	News	-0.043	Trump (57.6%)
Utah	11	Biden supporting	0.015	Trump (58.2%)
North Dakota	8	Trump Speech	0.204	Trump (65.5%)
Kansas	8	Trump Speech	0.202	Trump (56.2%)


TextBlob analysis:
mean polarity and mean subjectivity
(one tweet, one sentiment) vs (one user, one sentiment)
Trump
Biden
tweet level
user level
Average value
0.0
0.1
0.2
0.3
0.4
0.5
Polarity
Subjectivity
Feature