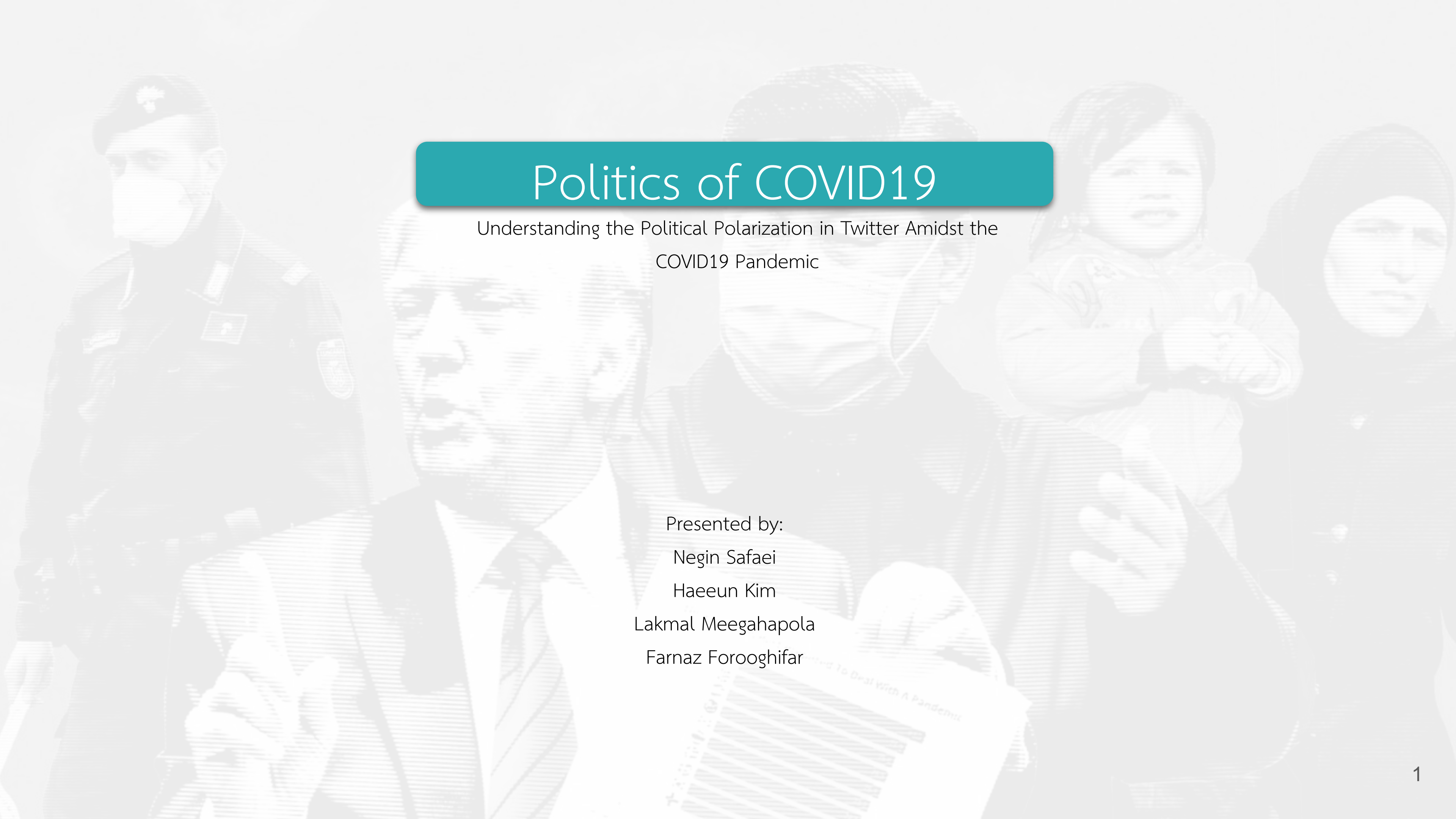


Politics of COVID19

Understanding the Political Polarization in Twitter Amidst the COVID19 Pandemic

Presented by:
Negin Safaei
Haeeun Kim
Lakmal Meegahapola
Farnaz Forooghifar

Outline

- Introduction
- Motivation
- Dataset
  - Data Collection
  - Feature Extraction
  - Available Data
- Descriptive Data Analysis
  - Metadata
  - Sentiment
  - Word and Hashtag
  - Latent Dirichlet Allocation (LDA)
- Statistical Analysis
- Binary Political Affinity Inference
- Discussion
- Conclusion and Future Work

Introduction

- **COVID19** pandemic

  Social media data (Twitter)

- COVID19 itself may be apolitical, but its’ influence on society is political.

It’s affected by

- Nationalities
- Geographies → **USA**
- Financial status
- Social roles
- Political affinities

Motivation

- **RQ1**: What trends and patterns can be observed with Twitter data regarding the political affinity in the USA during COVID19 pandemic?
- **RQ2**: Can binary political affinity be inferred using text and metadata of tweets?

Understand the most pressing concerns

Contents of social media → Bias perceptions

- **Goal:** Investigation the correlation between users' tweets and their political affinity and reactions

Project Overview and Task Distribution

Account selection → Data collection → Data analysis → Report and presentation

Account selection

Celebrities

Politicians

Media

Data collection

19th April- 31st May

200 tweets / day

Data analysis

LDA

Sentiment analysis

Hashtag analysis

Word analysis

...

Report and presentation

Dataset - Data Collection

- Tweet timestamps: April 5, 2020 to May 31, 2020 (collected between April 19 and May 31)
- Each day 200 tweets per user
- Each user's followers >= 25 k
- Categories:
  - Political affinities:
    - Democrat
    - Republican
  - Account types:
    - Politicians
    - Celebrities
    - Media
- 252 users in total

Category	Democrat	Republican
Politicians	50	49
Celebrities	49	45
Media	32	27

Dataset - Available Data

- After omitting duplicates we collected ~144000 tweets.
- Account activities

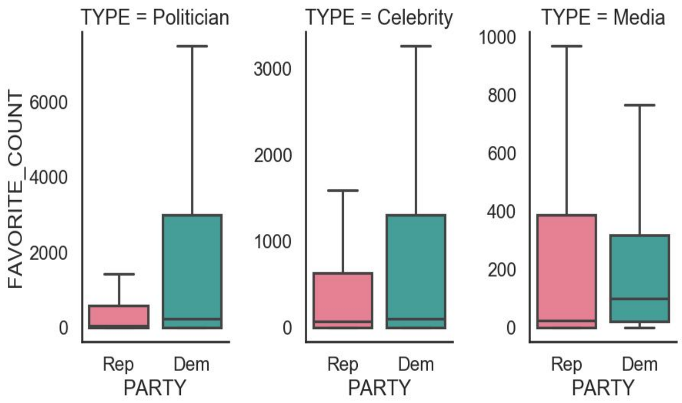
Category	Democrat	Republican
Politicians	17011	17420
Celebrities	13413	**23610**
Media	**47974**	**24450**

Dataset - Feature Extraction

Feature Group	Definition
META	Tweet specific features captured using the API
LDA	Topic modeling (3 <= α <= 150)
SENTI	VADER: positive, negative, neutral, and compound scores
	NLTK: a sentiment score between -1 and +1 for each sentence
HASH	10 most common hashtags for each category
WORD	10 most common words for each category

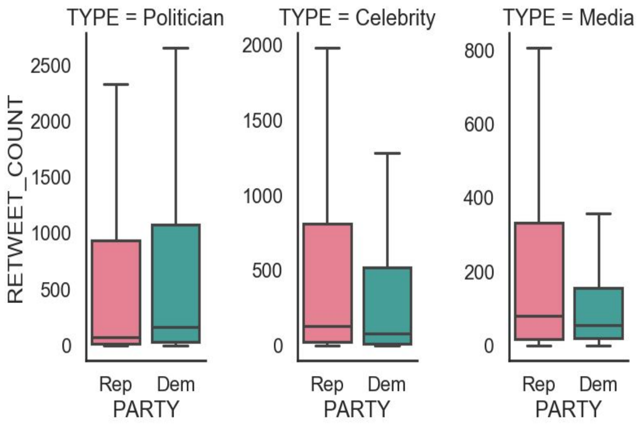
Descriptive Data Analysis

- Rep: Politicians = Celebrity = Media
- Dem: Politician > Celebrity > Media



Favorite count per category

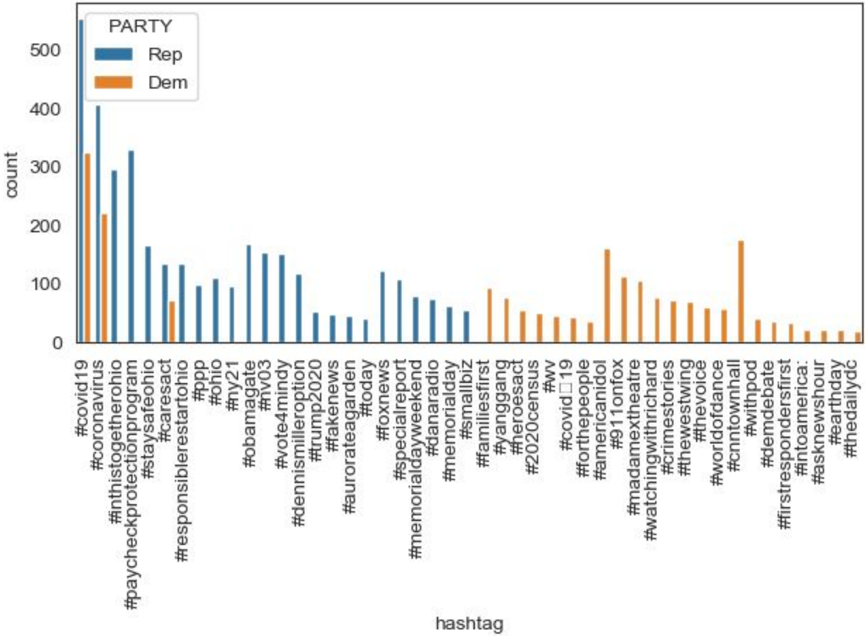
- Rep: Politicians > Celebrity > Media
- Dem: Politician > Celebrity > Media



Retweet count per category

Descriptive Data Analysis

- Hashtag Analysis

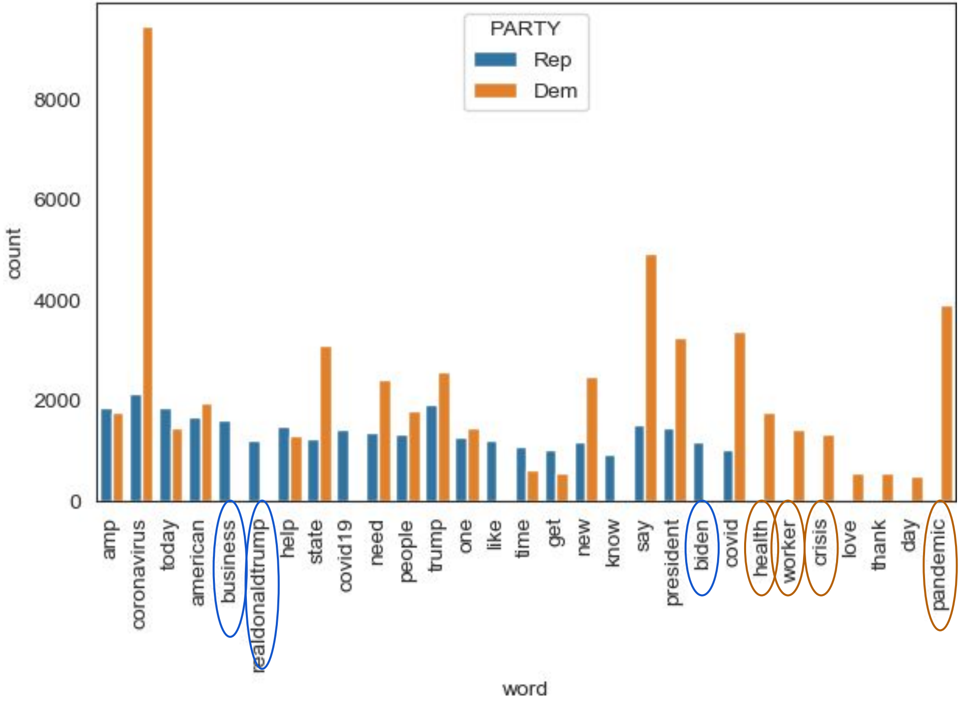


Five most-used **hashtags** for each affinity

	Democrats	Republicans
1	covid19	covid19
2	coronavirus	coronavirus
3	cnntownhall	paycheckprotectionprogram
4	americanidol	inthistogetherohio
5	familiesfirst	obamagate

Descriptive Data Analysis

- Word Analysis

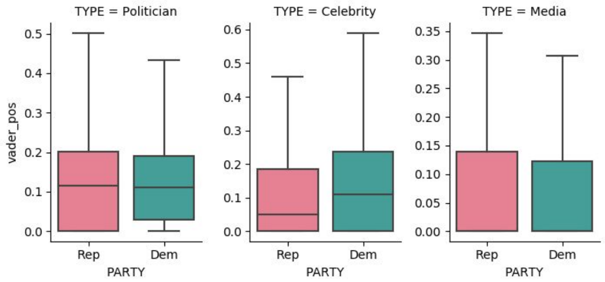


Five most-used **words** for each affinity

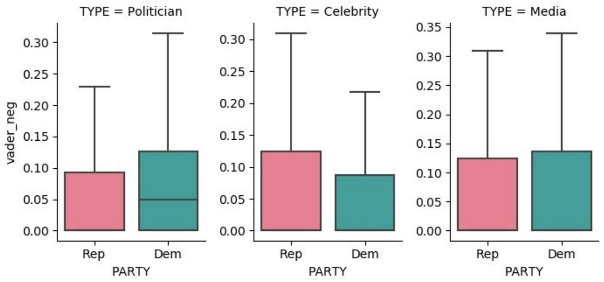
	Democrats	Republicans
1	coronavirus	coronavirus
2	say	trump
3	pandemic	today
4	covid	american
5	president	business

Descriptive Data Analysis

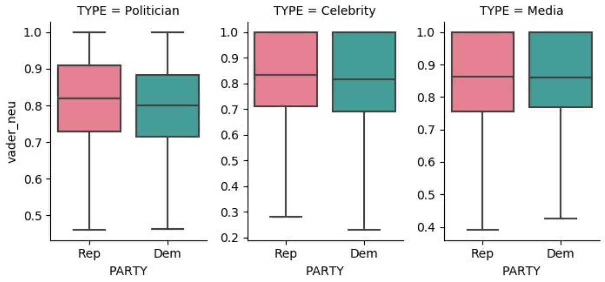
- Sentiment Analysis (Vader)



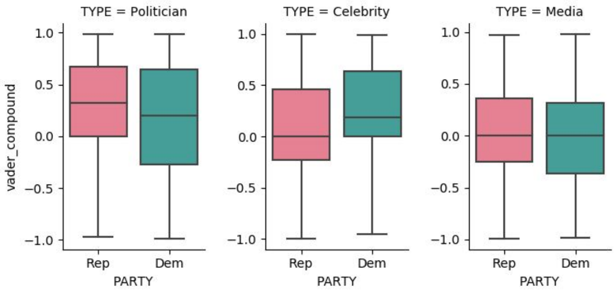
Positive sentiment score



Negative sentiment score



Neutral sentiment score



Compound sentiment score

Descriptive Data Analysis

- Topic Modelling

Words							Topic Name
coronavirus	pandemic	amid	lockdown	say	world	crisis	Global lockdown
business	small	program	loan	protection	help	relief	Help small business
trump	medium	twitter	president	fact	tweet	news	News about president Trump
story	great	news	good	show	new	today	Optimistic news
new	york	coronavirus	city	state	covid	say	New York situation
biden	joe	trump	flynn	president	former	say	Democrat vs republican
back	get	work	people	american	need	going	Get back to work for economy
worker	care	health	line	front	essential	pandemic	essential health care in pandemic

Statistical Analysis

- Objective: to identify features that would allow to discriminate the two political affinities.
- Metrics: t-statistic, p-value, Cohen's-d with 95% confidence interval
- Higher the t-statistic and Cohen's-d, higher the possibility of discriminating between the affinities using the feature.
- To interpret Cohen's-d, we use the widely accepted rule-of-thumb
  - 0.2 - small effect size
  - 0.5 - medium effect size
  - 0.8 - large effect size
- A confidence interval that does not include zero increases the confidence in the calculated effect size.

Statistical Analysis

Table 2: Comparative statistics of features across classes "democrats" and "republicans": t-statistic, p-value, and cohen's-d with 95% confidence intervals. Features are grouped by feature group and sorted based on the decreasing order of cohen's-d. Only five features each from HASH and WORD feature groups are presented here due to space limitations.

		Democrats		Republicans		Statistics			
Feature	Feature Group	mean	std	mean	std	abs. mean diff.	t-statistic	p-value	cohen's-d [95% CI]
number of urls	META	0.6670	0.6670	0.4843	0.5252	0.1872	67.72	$< 10^{-5}$	0.3576, [0.3472, 0.3681]
retweeted	META	0.0768	0.2664	0.1447	0.3518	0.0678	41.57	$< 10^{-5}$	0.2174, [0.2069, 0.2278]
retweet count	META	1236.43	8556.12	1590.86	5910.69	354.43	8.96	$< 10^{-5}$	0.0482, [0.0378, 0.0585]
favorite count	META	3473.60	31800.30	2350.83	13161.60	1122.77	8.45	$< 10^{-5}$	0.0461, [0.0357, 0.0565]
vader compound	SENTI	0.0578	0.4878	0.1030	0.4738	0.0451	17.69	$< 10^{-5}$	0.0938, [0.0834, 0.1042]
vader positive	SENTI	0.0962	0.1189	0.1062	0.1342	0.0100	14.99	$< 10^{-5}$	0.0789, [0.0685, 0.0893]
vader negative	SENTI	0.0738	0.1007	0.0665	0.1049	0.0073	13.42	$< 10^{-5}$	0.0709, [0.0605, 0.0813]
vader neutral	SENTI	0.8298	0.1391	0.8271	0.1535	0.0027	3.51	$< 10^{-3}$	0.0185, [0.0081, 0.0288]
nltk sentiment	SENTI	0.0638	0.1812	0.0672	0.2010	0.0033	3.35	$< 10^{-3}$	0.0176, [0.0073, 0.0280]
coronavirus	WORD	0.1506	0.3577	0.0678	0.2515	0.0827	51.91	$< 10^{-5}$	0.2677, [0.2577, 0.2777]
realdonaldtrump	WORD	0.0058	0.0763	0.0401	0.1962	0.0342	46.53	$< 10^{-5}$	0.2304, [0.2201, 0.2401]
pandemic	WORD	0.0709	0.2567	0.0258	0.1587	0.0450	40.79	$< 10^{-5}$	0.2112, [0.2012, 0.2212]
say	WORD	0.0820	0.2744	0.0390	0.1936	0.0430	35.11	$< 10^{-5}$	0.1810, [0.1711, 0.1910]
health	WORD	0.0559	0.2297	0.0309	0.1731	0.0249	23.84	$< 10^{-5}$	0.1226, [0.1156, 0.1297]
#paycheckprotectionprogram	HASH	0.0001	0.0084	0.0096	0.0972	0.0095	28.05	$< 10^{-5}$	0.1373, [0.1276, 0.1478]
#inthistogetherohio	HASH	0.0000	0.0034	0.0042	0.0647	0.0042	18.71	$< 10^{-5}$	0.0916, [0.0819, 0.1015]
#ppp	HASH	0.0001	0.0103	0.0036	0.0601	0.0035	16.63	$< 10^{-5}$	0.0815, [0.0716, 0.0918]
#coronavirus	HASH	0.0098	0.0988	0.0188	0.1361	0.0090	15.06	$< 10^{-5}$	0.0757, [0.0658, 0.0889]
#ohio	HASH	0.0000	0.0048	0.0028	0.0523	0.0027	14.96	$< 10^{-5}$	0.0732, [0.0631, 0.0839]

Statistical Analysis

Table 2: Comparative statistics of features across classes "democrats" and "republicans": t-statistic, p-value, and cohen's-d with 95% confidence intervals. Features are grouped by feature group and sorted based on the decreasing order of cohen's-d. Only five features each from HASH and WORD feature groups are presented here due to space limitations.

		Democrats		Republicans		Statistics			
Feature	Feature Group	mean	std	mean	std	abs. mean diff.	t-statistic	p-value	cohen's-d [95% CI]
number of urls	META	0.6670	0.6670	0.4843	0.5252	0.1872	67.72	$< 10^{-5}$	0.3576, [0.3472, 0.3681]
retweeted	META	0.0768	0.2664	0.1447	0.3518	0.0678	41.57	$< 10^{-5}$	0.2174, [0.2069, 0.2278]
retweet count	META	1236.43	8556.12	1590.86	5910.69	354.43	8.96	$< 10^{-5}$	0.0482, [0.0378, 0.0585]
favorite count	META	3473.60	31800.30	2350.83	13161.60	1122.77	8.45	$< 10^{-5}$	0.0461, [0.0357, 0.0565]
vader compound	SENTI	0.0578	0.4878	0.1030	0.4738	0.0451	17.69	$< 10^{-5}$	0.0938, [0.0834, 0.1042]
vader positive	SENTI	0.0962	0.1189	0.1062	0.1342	0.0100	14.99	$< 10^{-5}$	0.0789, [0.0685, 0.0893]
vader negative	SENTI	0.0738	0.1007	0.0665	0.1049	0.0073	13.42	$< 10^{-5}$	0.0709, [0.0605, 0.0813]
vader neutral	SENTI	0.8298	0.1391	0.8271	0.1535	0.0027	3.51	$< 10^{-3}$	0.0185, [0.0081, 0.0288]
nltk sentiment	SENTI	0.0638	0.1812	0.0672	0.2010	0.0033	3.35	$< 10^{-3}$	0.0176, [0.0073, 0.0280]
coronavirus	WORD	0.1506	0.3577	0.0678	0.2515	0.0827	51.91	$< 10^{-5}$	0.2677, [0.2577, 0.2777]
realdonaldtrump	WORD	0.0058	0.0763	0.0401	0.1962	0.0342	46.53	$< 10^{-5}$	0.2304, [0.2201, 0.2401]
pandemic	WORD	0.0709	0.2567	0.0258	0.1587	0.0450	40.79	$< 10^{-5}$	0.2112, [0.2012, 0.2212]
say	WORD	0.0820	0.2744	0.0390	0.1936	0.0430	35.11	$< 10^{-5}$	0.1810, [0.1711, 0.1910]
health	WORD	0.0559	0.2297	0.0309	0.1731	0.0249	23.84	$< 10^{-5}$	0.1226, [0.1156, 0.1297]
#paycheckprotectionprogram	HASH	0.0001	0.0084	0.0096	0.0972	0.0095	28.05	$< 10^{-5}$	0.1373, [0.1276, 0.1478]
#inthistogetherohio	HASH	0.0000	0.0034	0.0042	0.0647	0.0042	18.71	$< 10^{-5}$	0.0916, [0.0819, 0.1015]
#ppp	HASH	0.0001	0.0103	0.0036	0.0601	0.0035	16.63	$< 10^{-5}$	0.0815, [0.0716, 0.0918]
#coronavirus	HASH	0.0098	0.0988	0.0188	0.1361	0.0090	15.06	$< 10^{-5}$	0.0757, [0.0658, 0.0889]
#ohio	HASH	0.0000	0.0048	0.0028	0.0523	0.0027	14.96	$< 10^{-5}$	0.0732, [0.0631, 0.0839]

Statistical Analysis

Table 2: Comparative statistics of features across classes "democrats" and "republicans": t-statistic, p-value, and cohen's-d with 95% confidence intervals. Features are grouped by feature group and sorted based on the decreasing order of cohen's-d. Only five features each from HASH and WORD feature groups are presented here due to space limitations.

		Democrats		Republicans		Statistics			
Feature	Feature Group	mean	std	mean	std	abs. mean diff.	t-statistic	p-value	cohen's-d [95% CI]
number of urls	META	0.6670	0.6670	0.4843	0.5252	0.1872	67.72	$< 10^{-5}$	0.3576, [0.3472, 0.3681]
retweeted	META	0.0768	0.2664	0.1447	0.3518	0.0678	41.57	$< 10^{-5}$	0.2174, [0.2069, 0.2278]
retweet count	META	1236.43	8556.12	1590.86	5910.69	354.43	8.96	$< 10^{-5}$	0.0482, [0.0378, 0.0585]
favorite count	META	3473.60	31800.30	2350.83	13161.60	1122.77	8.45	$< 10^{-5}$	0.0461, [0.0357, 0.0565]
vader compound	SENTI	0.0578	0.4878	0.1030	0.4738	0.0451	17.69	$< 10^{-5}$	0.0938, [0.0834, 0.1042]
vader positive	SENTI	0.0962	0.1189	0.1062	0.1342	0.0100	14.99	$< 10^{-5}$	0.0789, [0.0685, 0.0893]
vader negative	SENTI	0.0738	0.1007	0.0665	0.1049	0.0073	13.42	$< 10^{-5}$	0.0709, [0.0605, 0.0813]
vader neutral	SENTI	0.8298	0.1391	0.8271	0.1535	0.0027	3.51	$< 10^{-3}$	0.0185, [0.0081, 0.0288]
nltk sentiment	SENTI	0.0638	0.1812	0.0672	0.2010	0.0033	3.35	$< 10^{-3}$	0.0176, [0.0073, 0.0280]
coronavirus	WORD	0.1506	0.3577	0.0678	0.2515	0.0827	51.91	$< 10^{-5}$	0.2677, [0.2577, 0.2777]
realdonaldtrump	WORD	0.0058	0.0763	0.0401	0.1962	0.0342	46.53	$< 10^{-5}$	0.2304, [0.2201, 0.2401]
pandemic	WORD	0.0709	0.2567	0.0258	0.1587	0.0450	40.79	$< 10^{-5}$	0.2112, [0.2012, 0.2212]
say	WORD	0.0820	0.2744	0.0390	0.1936	0.0430	35.11	$< 10^{-5}$	0.1810, [0.1711, 0.1910]
health	WORD	0.0559	0.2297	0.0309	0.1731	0.0249	23.84	$< 10^{-5}$	0.1226, [0.1156, 0.1297]
#paycheckprotectionprogram	HASH	0.0001	0.0084	0.0096	0.0972	0.0095	28.05	$< 10^{-5}$	0.1373, [0.1276, 0.1478]
#inthistogetherohio	HASH	0.0000	0.0034	0.0042	0.0647	0.0042	18.71	$< 10^{-5}$	0.0916, [0.0819, 0.1015]
#ppp	HASH	0.0001	0.0103	0.0036	0.0601	0.0035	16.63	$< 10^{-5}$	0.0815, [0.0716, 0.0918]
#coronavirus	HASH	0.0098	0.0988	0.0188	0.1361	0.0090	15.06	$< 10^{-5}$	0.0757, [0.0658, 0.0889]
#ohio	HASH	0.0000	0.0048	0.0028	0.0523	0.0027	14.96	$< 10^{-5}$	0.0732, [0.0631, 0.0839]

Statistical Analysis

Table 2: Comparative statistics of features across classes "democrats" and "republicans": t-statistic, p-value, and cohen's-d with 95% confidence intervals. Features are grouped by feature group and sorted based on the decreasing order of cohen's-d. Only five features each from HASH and WORD feature groups are presented here due to space limitations.

		Democrats		Republicans		Statistics			
Feature	**Feature Group**	**mean**	**std**	**mean**	**std**	**abs. mean diff.**	**t-statistic**	**p-value**	**cohen's-d [95% CI]**
number of urls	META	0.6670	0.6670	0.4843	0.5252	0.1872	67.72	$< 10^{-5}$	0.3576, [0.3472, 0.3681]
retweeted	META	0.0768	0.2664	0.1447	0.3518	0.0678	41.57	$< 10^{-5}$	0.2174, [0.2069, 0.2278]
retweet count	META	1236.43	8556.12	1590.86	5910.69	354.43	8.96	$< 10^{-5}$	0.0482, [0.0378, 0.0585]
favorite count	META	3473.60	31800.30	2350.83	13161.60	1122.77	8.45	$< 10^{-5}$	0.0461, [0.0357, 0.0565]
vader compound	SENTI	0.0578	0.4878	0.1030	0.4738	0.0451	17.69	$< 10^{-5}$	0.0938, [0.0834, 0.1042]
vader positive	SENTI	0.0962	0.1189	0.1062	0.1342	0.0100	14.99	$< 10^{-5}$	0.0789, [0.0685, 0.0893]
vader negative	SENTI	0.0738	0.1007	0.0665	0.1049	0.0073	13.42	$< 10^{-5}$	0.0709, [0.0605, 0.0813]
vader neutral	SENTI	0.8298	0.1391	0.8271	0.1535	0.0027	3.51	$< 10^{-3}$	0.0185, [0.0081, 0.0288]
nltk sentiment	SENTI	0.0638	0.1812	0.0672	0.2010	0.0033	3.35	$< 10^{-3}$	0.0176, [0.0073, 0.0280]
coronavirus	WORD	0.1506	0.3577	0.0678	0.2515	0.0827	51.91	$< 10^{-5}$	0.2677, [0.2577, 0.2777]
realdonaldtrump	WORD	0.0058	0.0763	0.0401	0.1962	0.0342	46.53	$< 10^{-5}$	0.2304, [0.2201, 0.2401]
pandemic	WORD	0.0709	0.2567	0.0258	0.1587	0.0450	40.79	$< 10^{-5}$	0.2112, [0.2012, 0.2212]
say	WORD	0.0820	0.2744	0.0390	0.1936	0.0430	35.11	$< 10^{-5}$	0.1810, [0.1711, 0.1910]
health	WORD	0.0559	0.2297	0.0309	0.1731	0.0249	23.84	$< 10^{-5}$	0.1226, [0.1156, 0.1297]
#paycheckprotectionprogram	HASH	0.0001	0.0084	0.0096	0.0972	0.0095	28.05	$< 10^{-5}$	0.1373, [0.1276, 0.1478]
#inthistogetherohio	HASH	0.0000	0.0034	0.0042	0.0647	0.0042	18.71	$< 10^{-5}$	0.0916, [0.0819, 0.1015]
#ppp	HASH	0.0001	0.0103	0.0036	0.0601	0.0035	16.63	$< 10^{-5}$	0.0815, [0.0716, 0.0918]
#coronavirus	HASH	0.0098	0.0988	0.0188	0.1361	0.0090	15.06	$< 10^{-5}$	0.0757, [0.0658, 0.0889]
#ohio	HASH	0.0000	0.0048	0.0028	0.0523	0.0027	14.96	$< 10^{-5}$	0.0732, [0.0631, 0.0839]

Statistical Analysis

- Results from the statistical analysis of topic model features across classes "democrats" and "republicans" for different number of topics

Table 3: Comparative statistics of topic model features across classes "democrats" and "republicans" for different number of topics. Mean and Maximum of t-statistic and cohen's-d based results are included.

# of topics	maximum t-statistic	mean t-statistic	maximum cohen's-d	mean cohen's-d
3	44.8921	29.8019	0.2390	0.1581
4	42.7796	21.2210	0.2283	0.1127
5	38.8133	18.8462	0.2074	0.1001
6	36.9744	20.8717	0.1977	0.1107
7	34.7234	19.1773	0.1856	0.0928
8	34.6744	17.5067	0.1857	0.0928
10	32.0082	15.6414	0.1715	0.0829
15	28.2354	13.6842	0.1493	0.0723
20	28.4012	12.2689	0.1484	0.0651
25	30.9418	11.5765	0.1668	0.0651
30	28.0685	10.8192	0.1500	0.0574

Inference

- We model a task to infer **binary political affinity** using different feature group combinations.
- We chose Random Forest Classifiers, considering the tabular nature of the dataset.
- Used feature groups are:
  - WORD - 10 words with highest t-statistic
  - HASH - 10 hashtags with highest t-statistic
  - META - 4 metadata from tweets including favorites, retweets, etc.
  - SENTI - 5 features derived from the sentiment analysis
  - LDA - 30 features derived from latent dirichlet allocation.
- Used leave k-participants out strategy to make sure that training and testing sets do not have data from the same user to avoid biases.
- Experiments repeated for five iterations using different

Inference

- Democrat vs. Republican inference results

Table 4: Inference Accuracies for Different Feature Group Combinations

Feature Group	Accuracy	Precision	Recall
Baseline	50.00%	50.00%	50.00%
HASH	64.36%	65.41%	61.59%
LDA	68.43%	68.83%	68.24%
WORD	69.41%	67.31%	69.32%
META	70.90%	70.99%	70.89%
SENTI	70.19%	70.43%	70.24%
META+HASH+WORD	76.23%	77.41%	76.28%
META+HASH+WORD+SENTI	79.37%	79.38%	79.49%
META+HASH+WORD+LDA	80.19%	80.21%	80.19%
META+HASH+WORD+LDA+SENTI	82.73%	82.71%	81.18%
META+HASH+WORD+LDA+SENTI+Type	83.78%	83.84%	83.78%

Inference

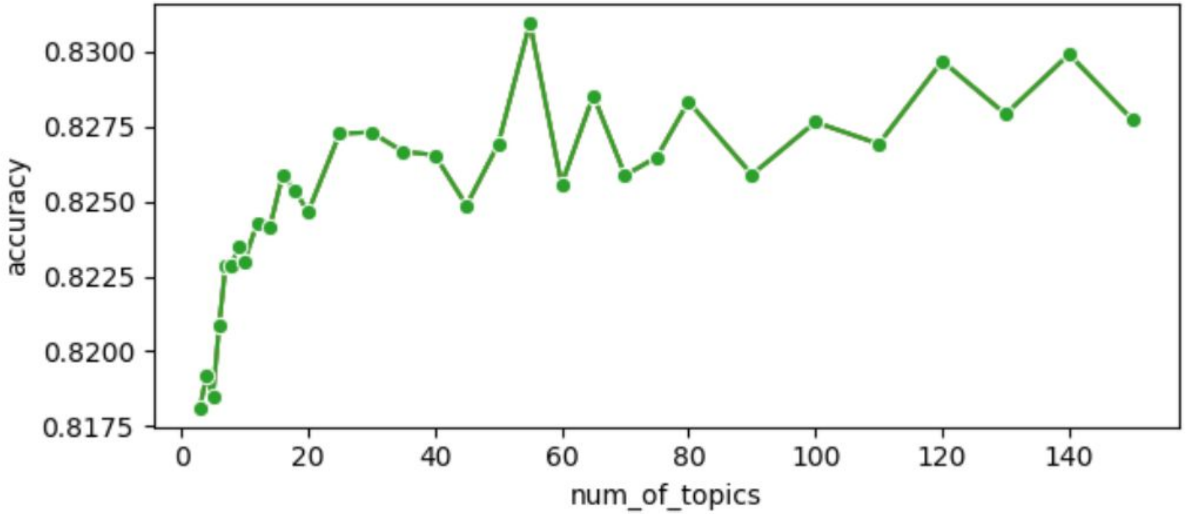


Figure 6: Number of topics vs. Average Accuracy using Random Forest Classifiers.

Summary

- Contribution 1: Discovering the different tweeting patterns between different political affinities
  - Most of the hashtags were used only by one political affinity.
  - Different topics of interest were revealed in the word analysis
  - Neutral sentiments were dominant in all categories and political affinity.

- Contribution 2: Inferring the political affinities from the tweet data.
  - Up to 82.73% of accuracy.
  - Various feature groups were combined.

Discussion

- Political polarization is observed even during the ***pandemic crisis***!
  - This extends the previous studies on political polarization in social media during the ***presidential election.***
- Question 1: ***Is disaster - or other seemingly apolitical events - really apolitical?***
- Question 2: ***What should (not) be done by the data scientists and computational social scientists?***
- Our machine-mediated inference model can be used to mitigate the polarization in social media.
  - ...which is especially important in critical situations like disaster and public health control.

Limitations

- Limited number of Twitter profiles
  - No more than 50 for each group, whose representativeness was assumed.
- Binary distinction of political affinity
  - In real-world, it always lies on a continuous spectrum.

⟶ Improvements can be made by having a detailed, concrete ground-truth dataset.

⟶ This requires empirical survey on the U.S. Twitter networks.

Future Work

- In-depth analysis on tweet content
  - Other NLP techniques can be applied.
  - *What kind of message was delivered by using different words and hashtags?*
- Temporal analysis
  - Align the tweet analysis to the major events and news updates during the given time period
  - *How immediately did each political affinity group react to different issues?*
- Other sectors and issues of the U.S. society
  - *Is such political polarization found in other (seemingly apolitical) sectors as well?*
- Other societies with different political landscapes
  - *Is such political polarization found in liberal/authoritarian/socialist societies as well?*

Thank You!

Any Questions?

Appendix

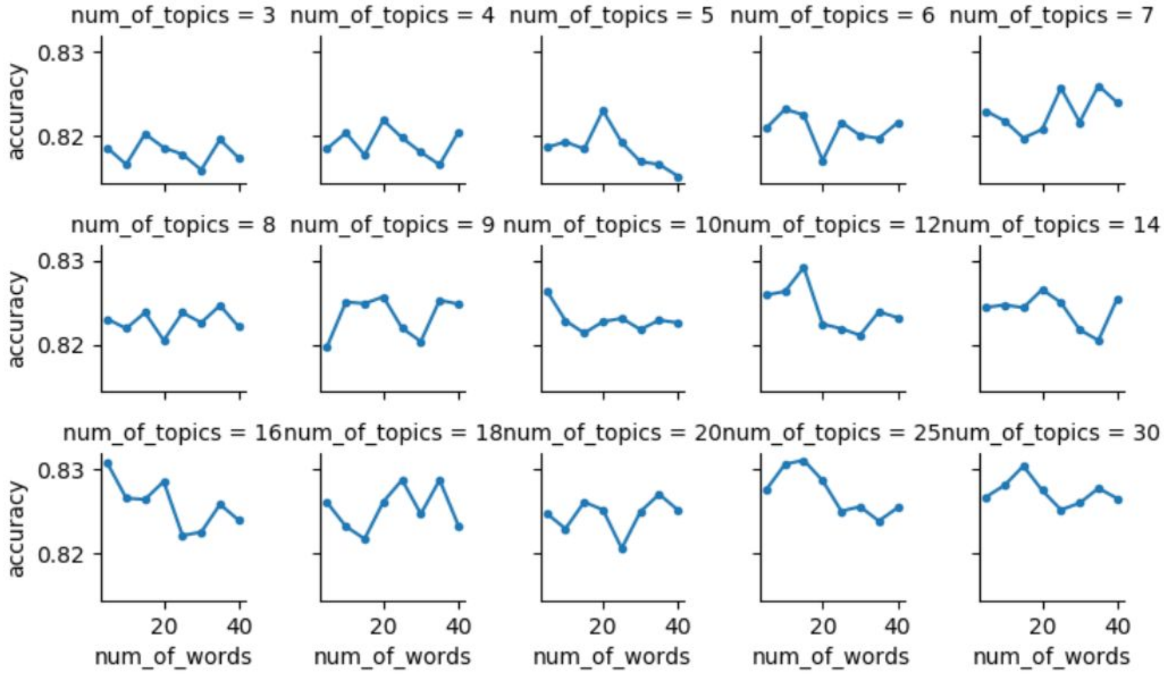


Figure 5: Number of words vs. Accuracy plot for different number of topics.

Appendix - NLTK Sentiments

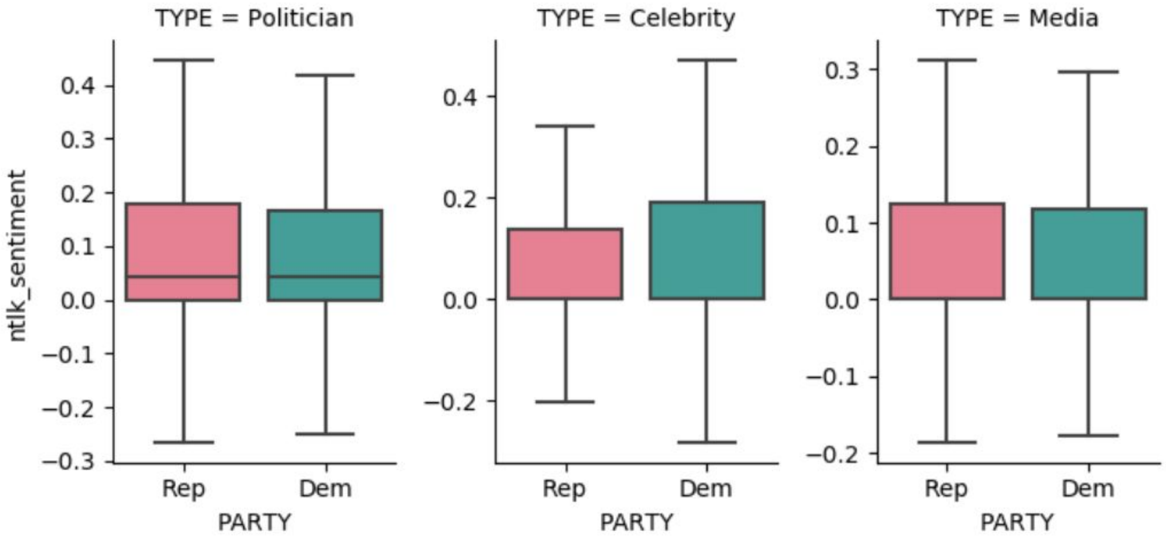
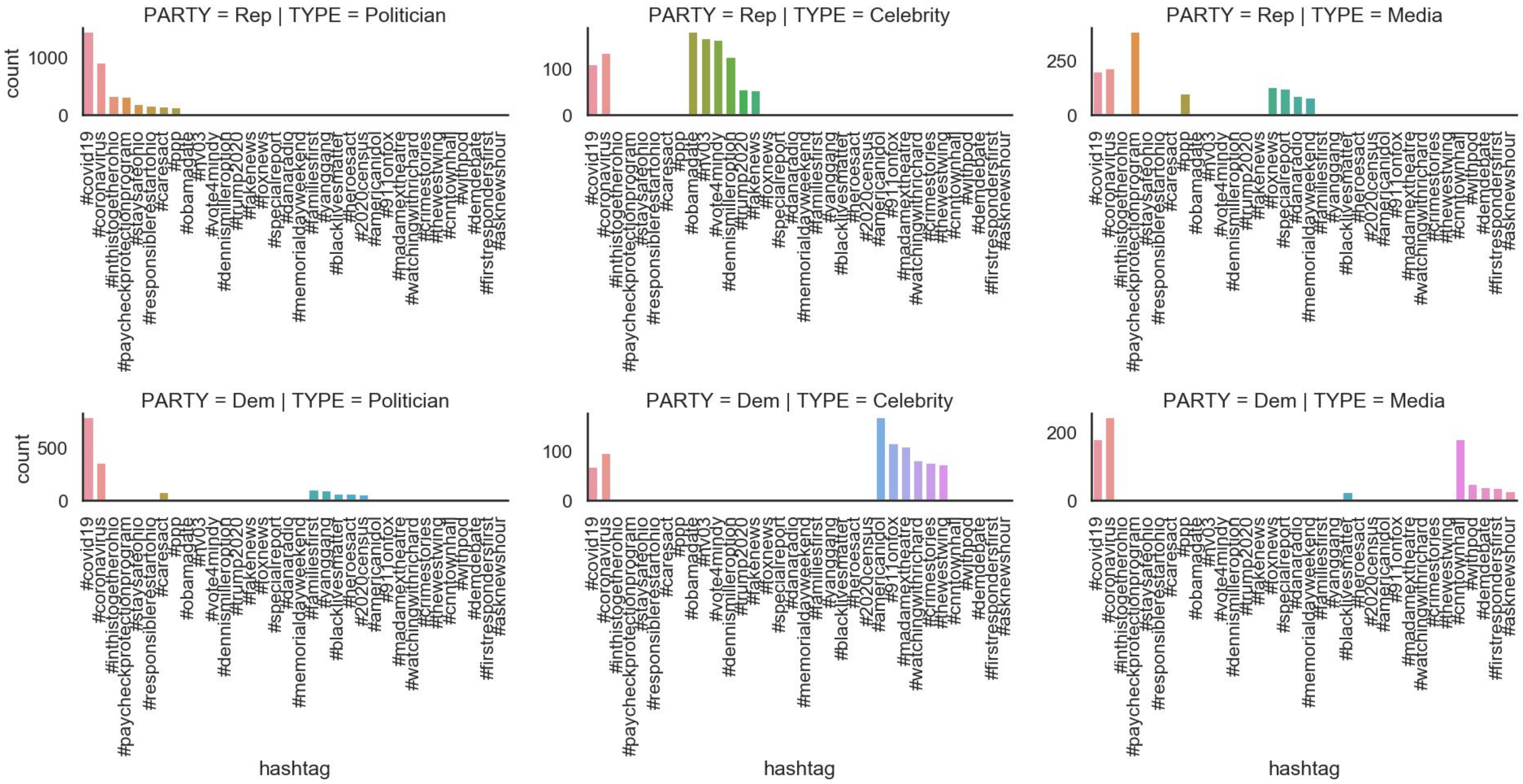


Figure 11: Boxplot for the feature NLTK sentiment

Appendix - Hashtag Analysis



Appendix - Word Analysis