Online Classifier Strategy

(2020 Presidential Campaign)

Election Campaign Team (June Update)

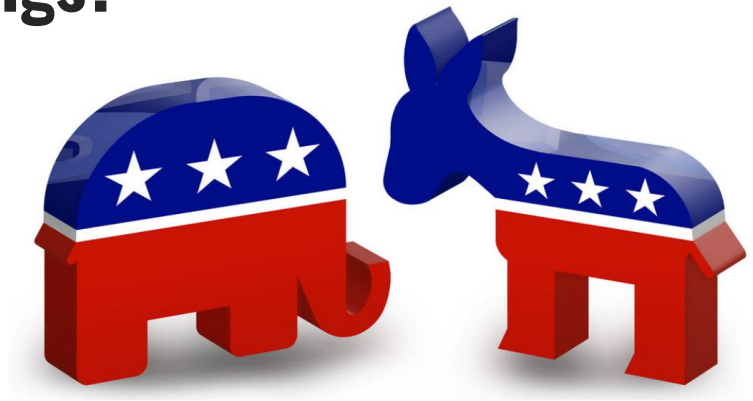
Melvin Lee, Evonne Tham, Lester Phong

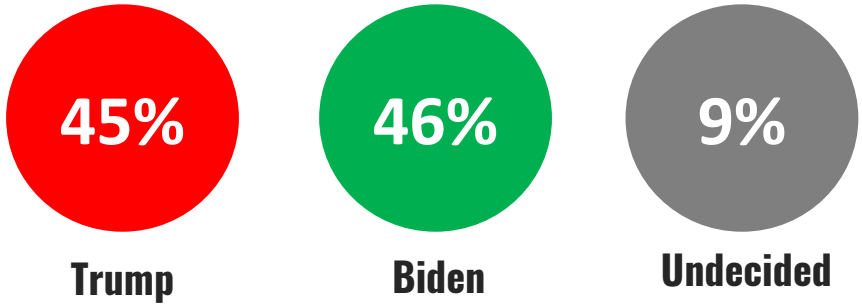
Background

BIDEN PRESIDENT

Where are we in terms of poll standings?

- Latest polls indicate a close race with marginal lead
- Biden lead has been diminishing and volatile
- Swing group and Trump supporters hold potential due to narrow margin
- Trump is very politically engaged online and most voters today are internet savvy





Source: *Trafalgar Group Michigan Polls*

Problem Statement

BIDEN PRESIDENT

Where are the gaps?

- Trump has a stronger and more effective presence in the digital space
- How can we optimize digital presence to gain more political supporters and insight?
- How can we structure online campaign for most impactful outreach?

How can it be addressed?

- Identify trending topics that voters are concerned about most
- Monitor trend towards democrat or republican over discussion platform
- Based on above, develop a repeatable classifier model to gauge online sentiment

Data Gathering

Most effective online source for baseline data?

1.7 billion users

330 million users

330 million users

Reddit Sub

1. r/democrats (133K)
2. r/republican (120K)

Considerations

- Twitter is short post based social media platform
- Facebook is sharing events and pictures
- Reddit primarily online bulletin with useful reservoir of content heavy discussion posts

Data Cleaning

BIDEN PRESIDENT

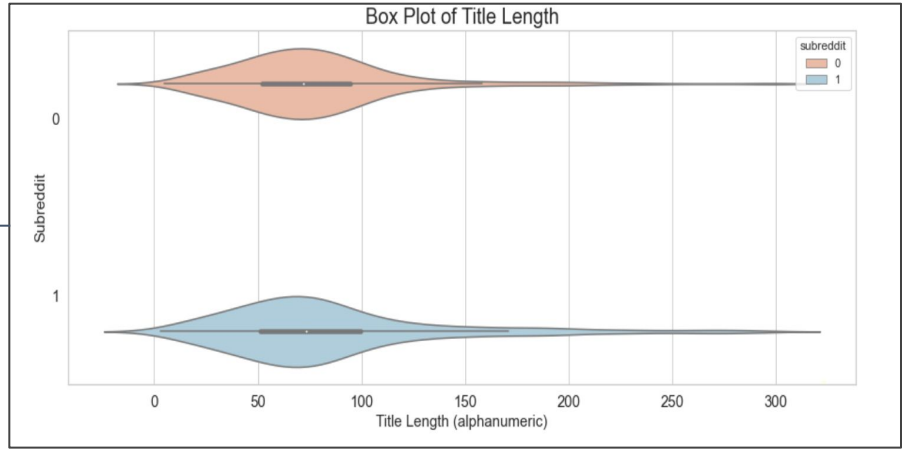
- Duplicate Posts
- Moderator Posts
- No Advertisements

1

Cleaning

2

EDA



3

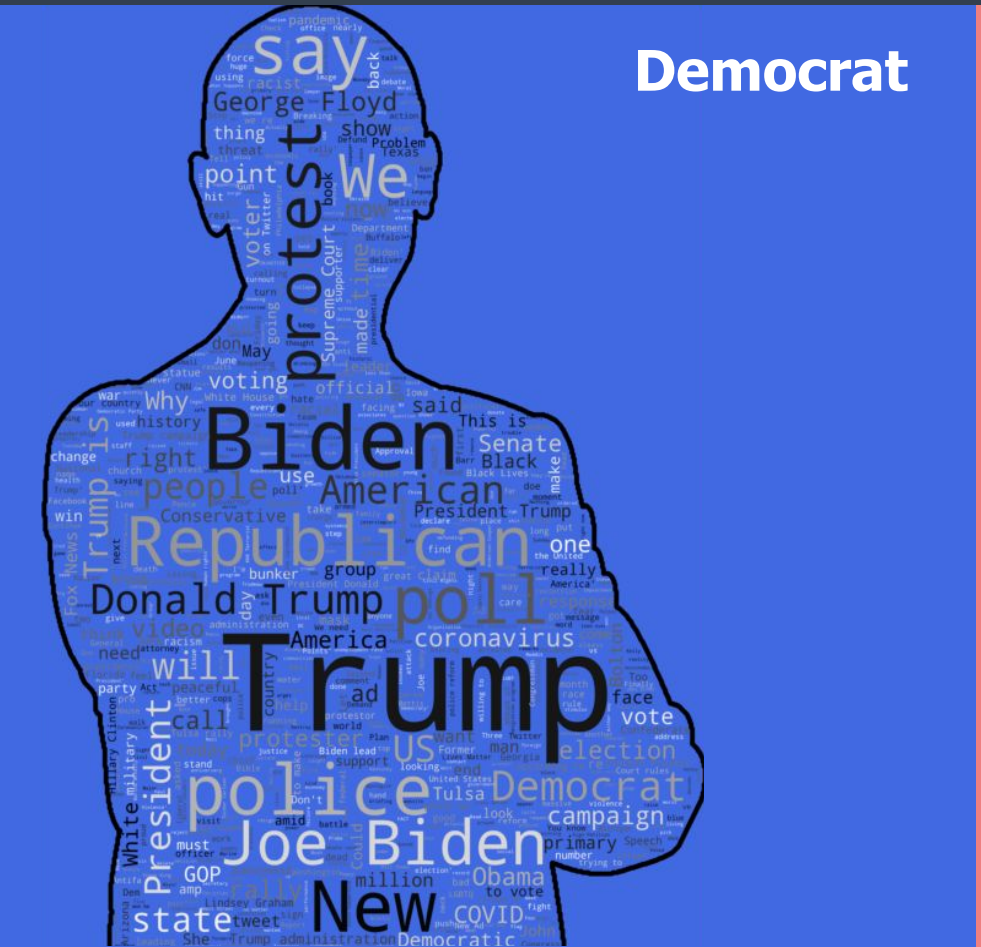
Pre-Processing

- URL prefixes
- Lemmatization
- Stop words
- Vectorization

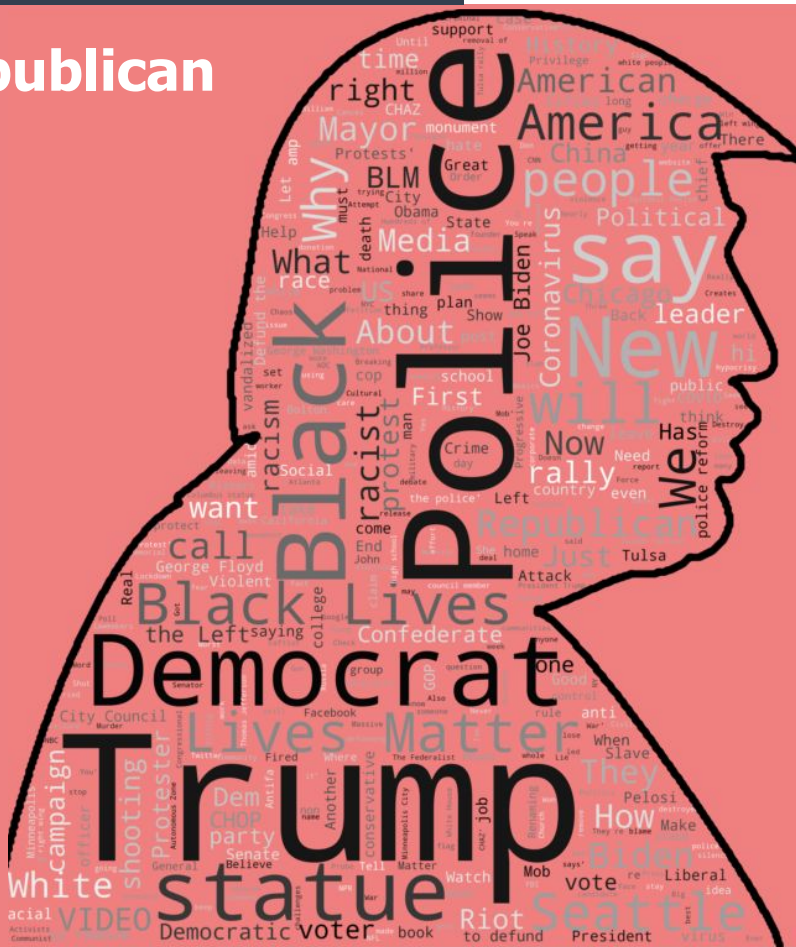
EDA - Visualisation

BIDEN PRESIDENT

Democrat

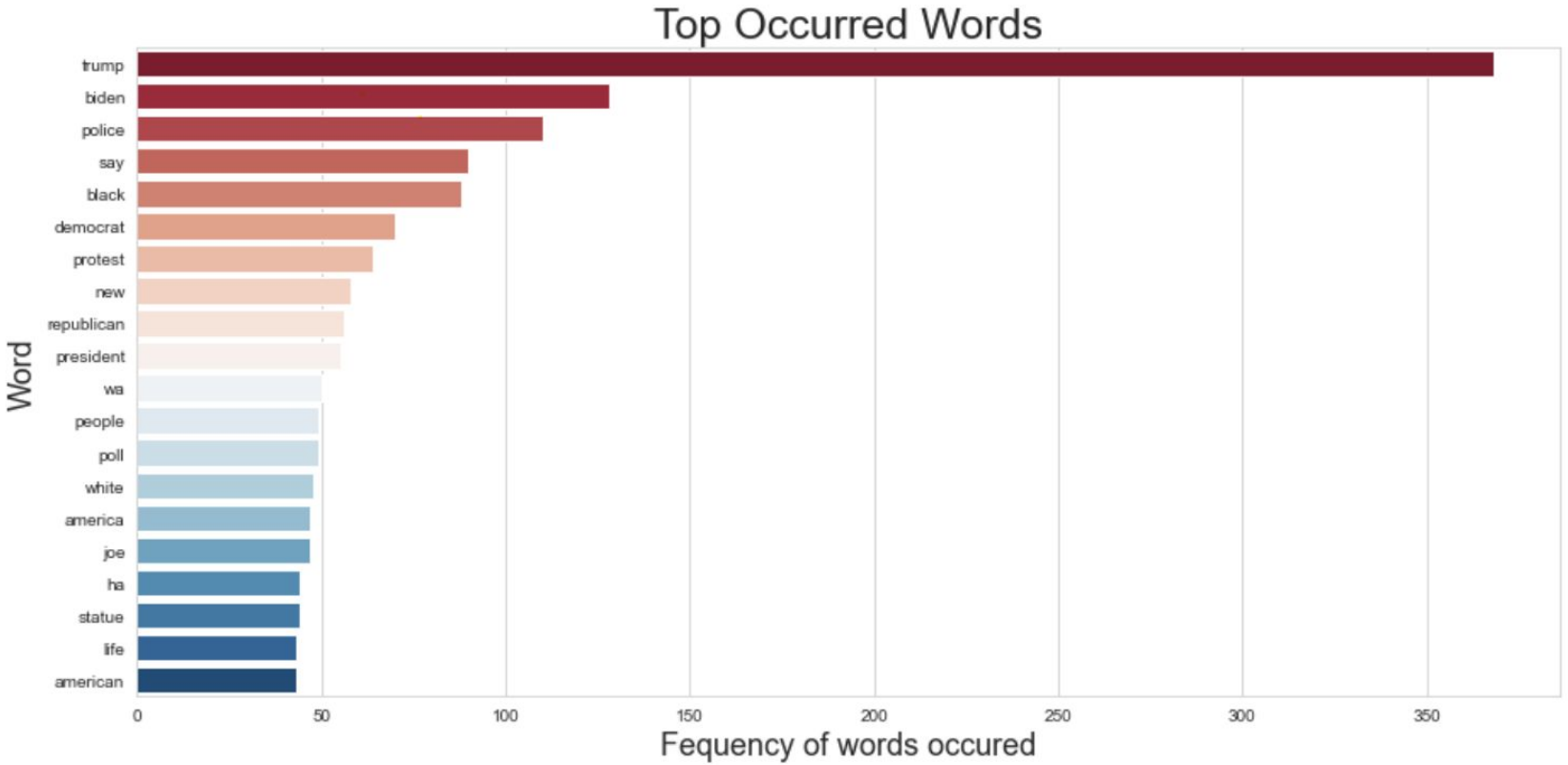


Republican



EDA - Visualization

Top Words from the Combined Production Set



Modeling

- Feature Extraction
  - CountVectorizer
  - TF-IDF *(term frequency-inverse document frequency)*
- Model Use
  - Linear Regression
  - Naive Bayes
    - Multinomial
- GridSearch

Model Combination	Best Score	Train Accuracy	Test Accuracy
Baseline Model	0.549383	NaN	NaN
CountVec_LR	0.719311	0.834581	0.708520
CountVec_NB	0.728293	0.886976	0.708520
TFID_LR	0.726048	0.889970	0.701794
TFID_NB	0.725299	0.894461	0.719731

	precision	recall	f1-score	support
Democrats	0.77	0.60	0.67	191
Republican	0.73	0.86	0.79	246
accuracy			0.75	437
macro avg	0.75	0.73	0.73	437
weighted avg	0.75	0.75	0.74	437

ROC Curve with AUC = 0.807

Best Parameters

- `cvec__max_df: 0.9`
- `cvec__max_features: 1500`
- `cvec__min_df: 2`
- `cvec__ngram_range: (1,1)`
- `cvec__stop_words: english`
- `nb__alpha: 1.5`

Findings

BIDEN PRESIDENT

Common Buzzwords

coronavirus

- New cases
- Death toll

black

- Protest
- Black Lives Matter
- Deaths

police

- Police Funding
- Policing Guidelines
- Reform Bill Scope

trump

- Central Figure
- Controversial
- Presidential Race

biden

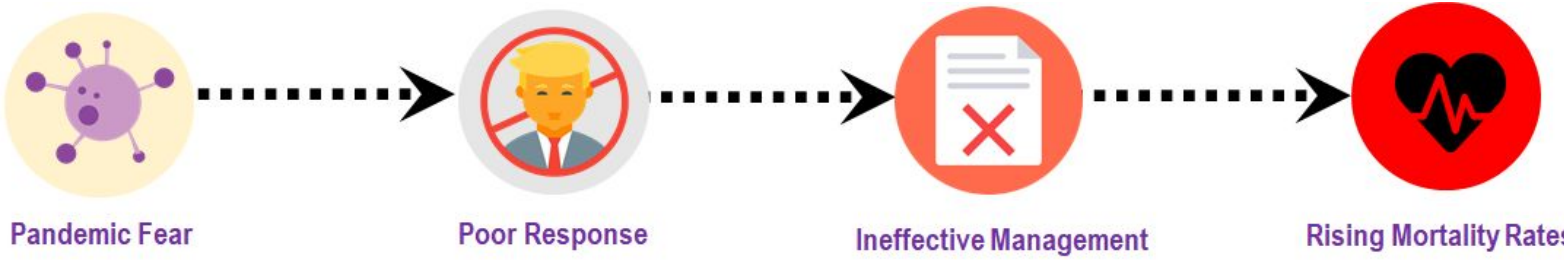
- Presidential Race
- Democratic icon

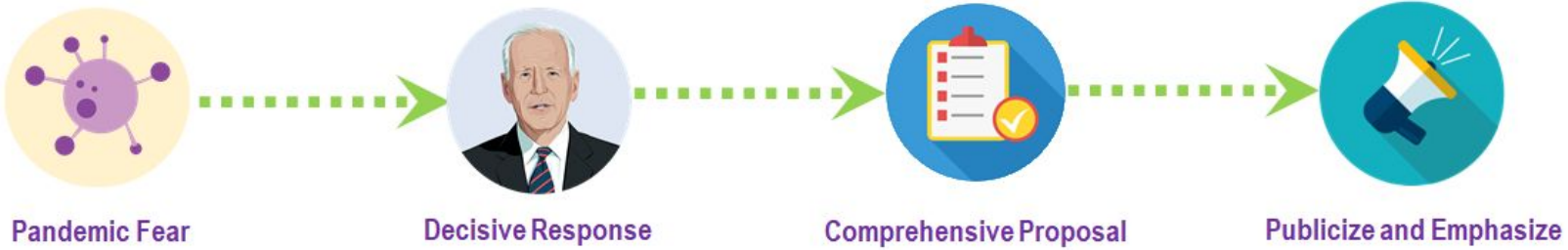
- Event recency heavily drives and impacts model
- Voter sentiment swings with current events
- Online mentions of Trump in reddit posts are more than double that of Biden's

Recommendations

BIDEN PRESIDENT

Campaign Opportunity #1





Recommendations

BIDEN PRESIDENT

Campaign Opportunity #2

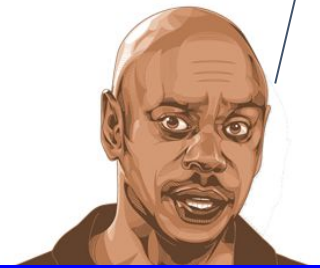
Solidarity

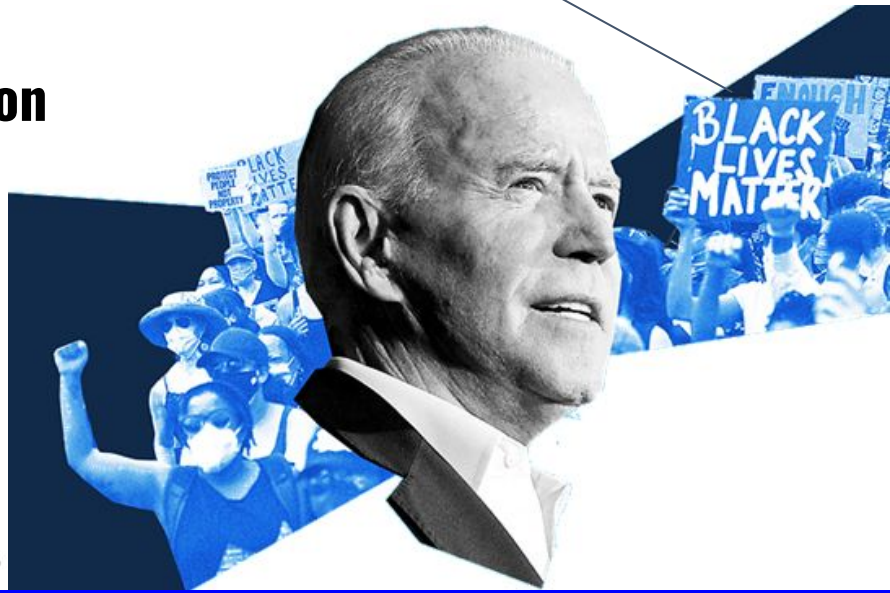
Inclusivity

Endorsement

Collaboration

Engagement





Recommendations

BIDEN PRESIDENT

Technical Recommendations

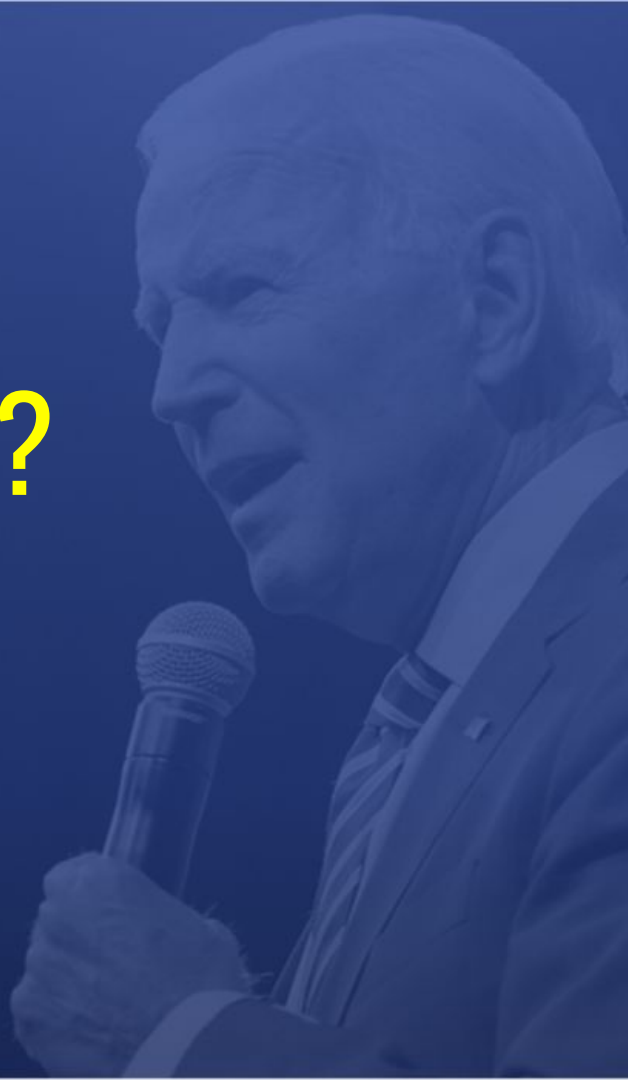
Additional Data Scrape

PUSHSHIFT

Wider API limits

Harness more information sources

Thank you, any questions?



BIDEN
PRESIDENT

Additional Support