The 2020 Presidential Race: NLP and Sentimental Analysis of Online News Articles

Caroline Sklaver, Wenyu Zeng, Bin Xing, Weining Hu

Outline

- Background
- Goals
- Data Collection & Pre-processing
- Term Frequency
- Topic Analysis
- Sentiment Analysis
- Modeling & Evaluation
- Conclusions

Background

- The 2020 United States presidential election is unprecedented in many ways in history
- As the U.S. is facing this pivotal election in the Fall, top news outlets serve as the major sources where most American voters get campaign news
- President Donald Trump is the leading Republican candidate seeking re-election
- Former Vice President Joe Biden is the leading Democratic candidate
- Bernie Sanders was a Democratic front-runner rivaling Joe Biden, but has since dropped out of the race
- The top US online newspapers include *The New York Times*, *The Guiardian*, *The Los Angeles Times*, *USA Today*, *The Wallstreet Journal*, *The Washington Post*, *Los Angeles Daily News*
- Previous research has been done regarding the media's influence in previous elections. We hope to extend these findings to the 2020 election.

Goals

Define key words or topics associated with each candidate

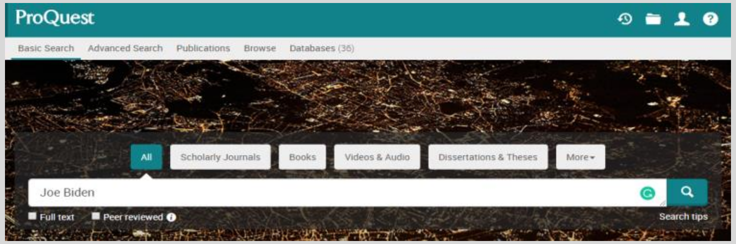
Term Frequency analysis	TF-IDF (term frequency-inverse document frequency)	Textacy (statement extraction)	Gensim (LDA topic modeling)

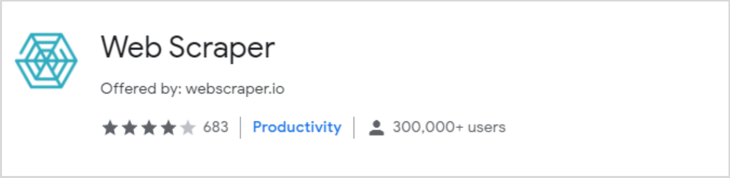
Build a model to predict the sentiment (positive, neutral, or negative) towards each candidate

Data Collection

	Joe Biden	Bernie Sanders	Donald Trump	Total
New York Times	150	150	150	450
Wall Street Journal	150	150	150	450
The Washington Post	150	150	150	450
Total News Articles	450	450	450	1350

Articles: Each presidential candidate has 450 articles, among which 150 for each news companies
Publication Date: January 1 – March 18, 2020
Scraping Keywords: Candidate full name + "Election" (i.e. "Joe Biden" AND "Election")
Database: ProQuest via GW Libraries
Scraping Tool: Web Scraper





Data Pre-processing & Annotation

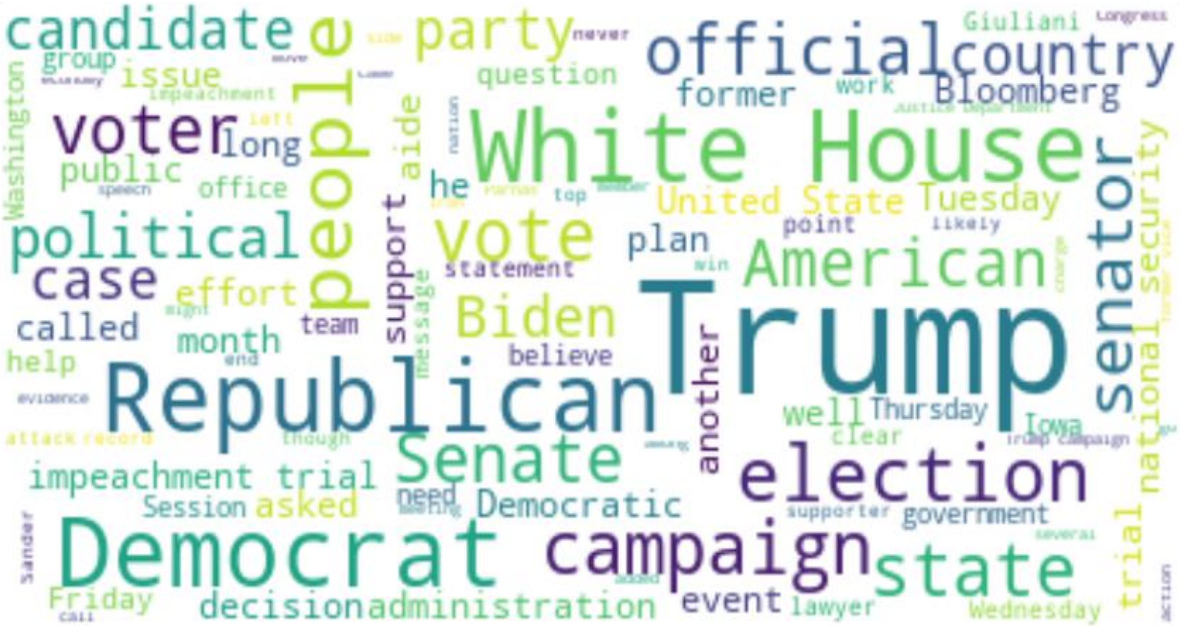
Data-Preprocessing: remove garbage words and non-alphabetic characters; lowercase and stem each word

Annotations: Identified 20 positive, 20 neutral, and 20 negative articles for each candidate

180 annotated articles out of 1350 total

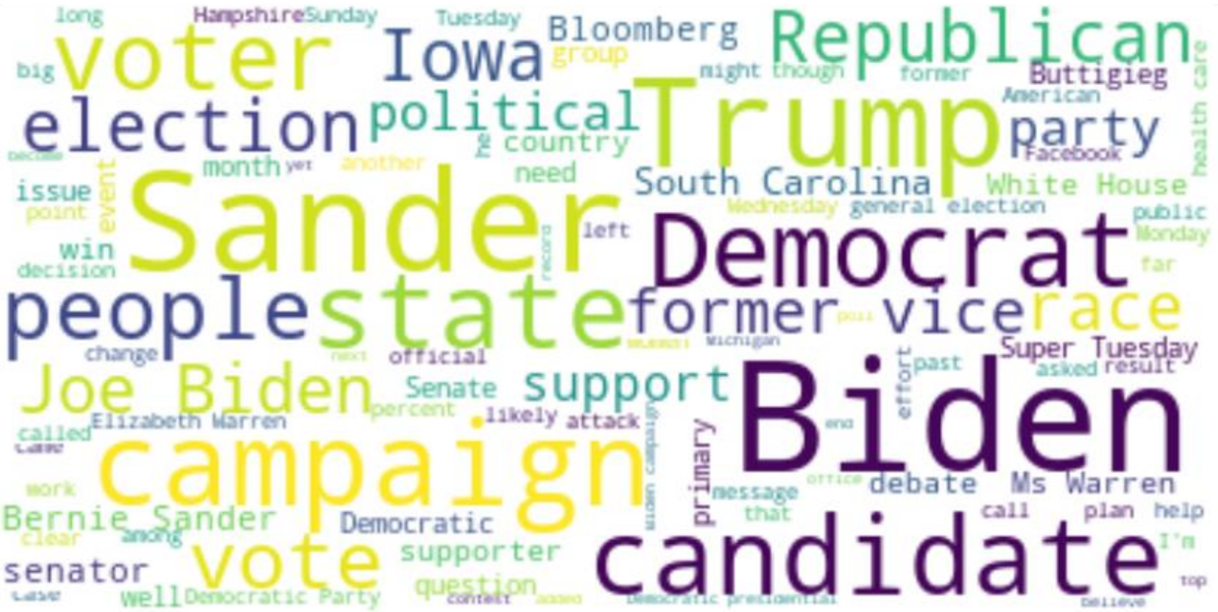
Only chose articles relevant to each candidate

Annotator used digression to determine sentiment of article (limitation)



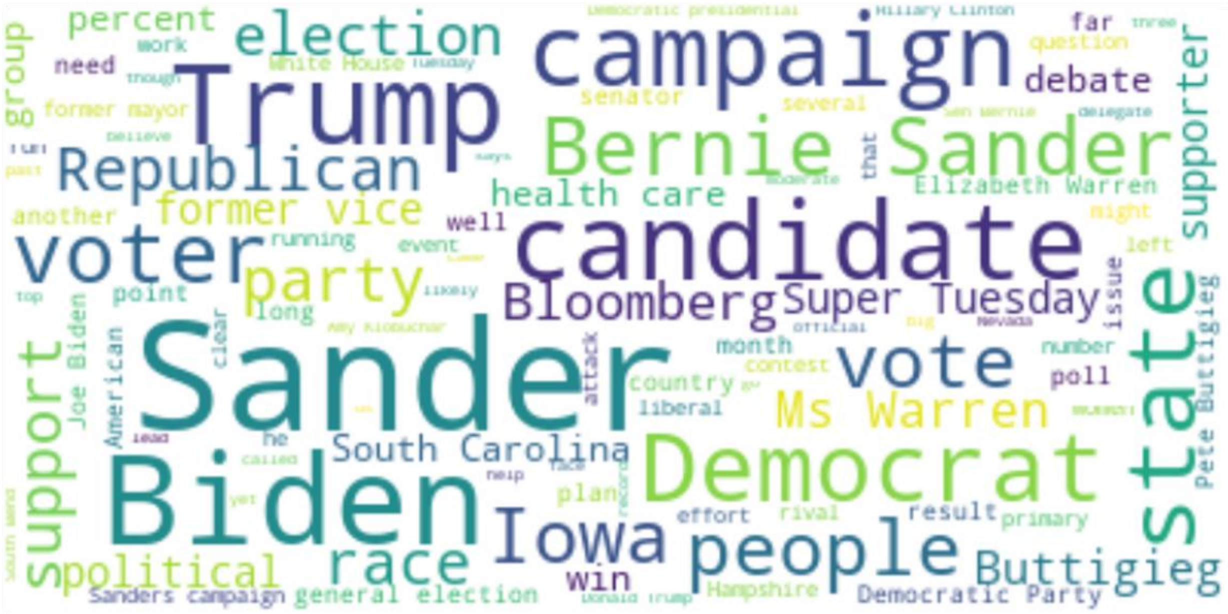
DONALD TRUMP

45TH U.S. PRESIDENT



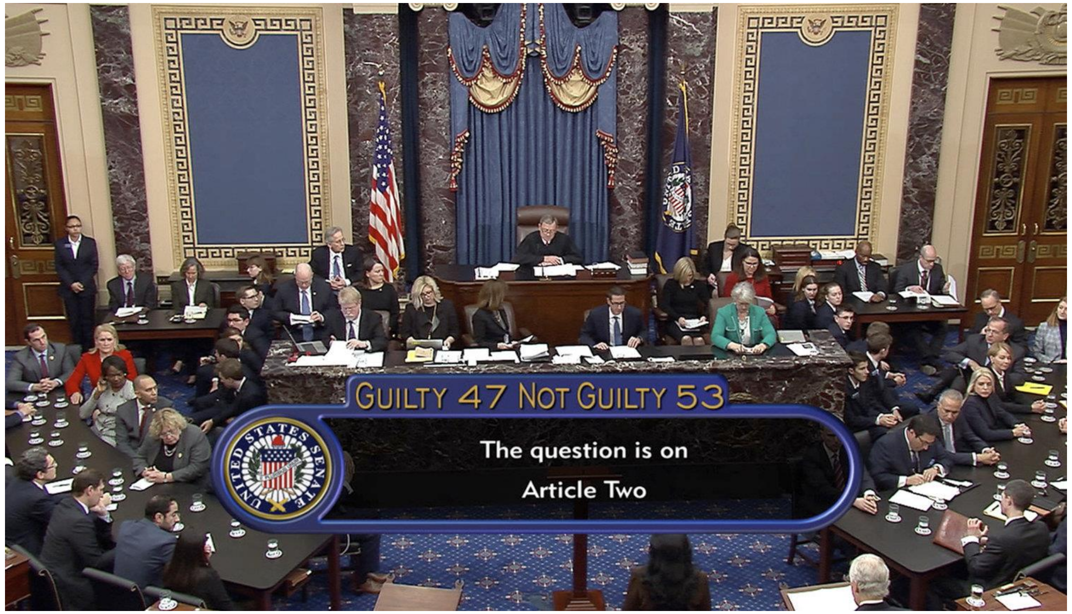
JOE BIDEN

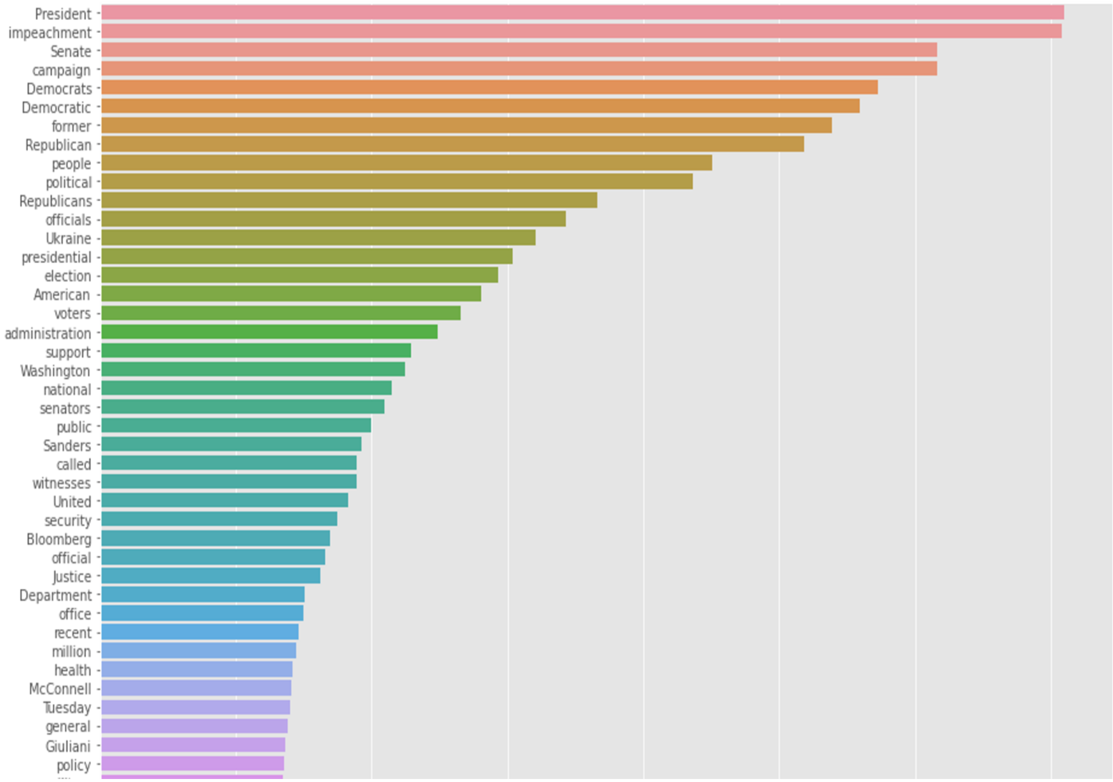
FORMER VICE PRESIDENT OF THE UNITED STATE



BERNIE SANDERS

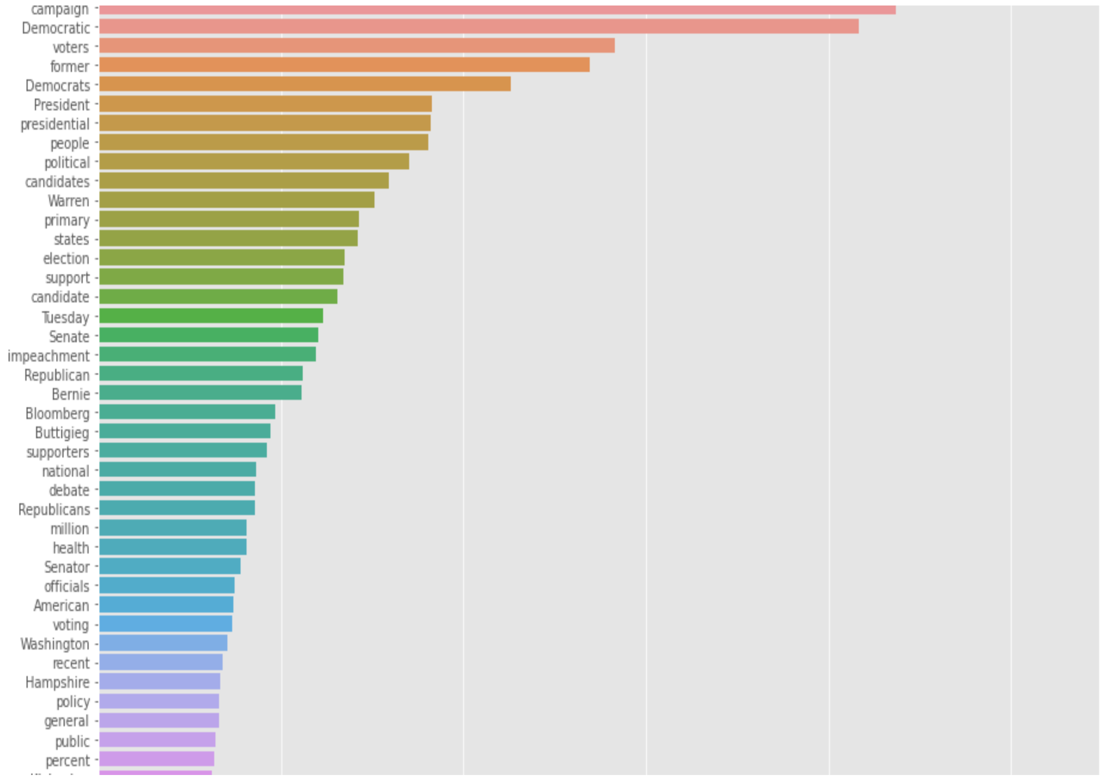
UNITED STATES SENATOR



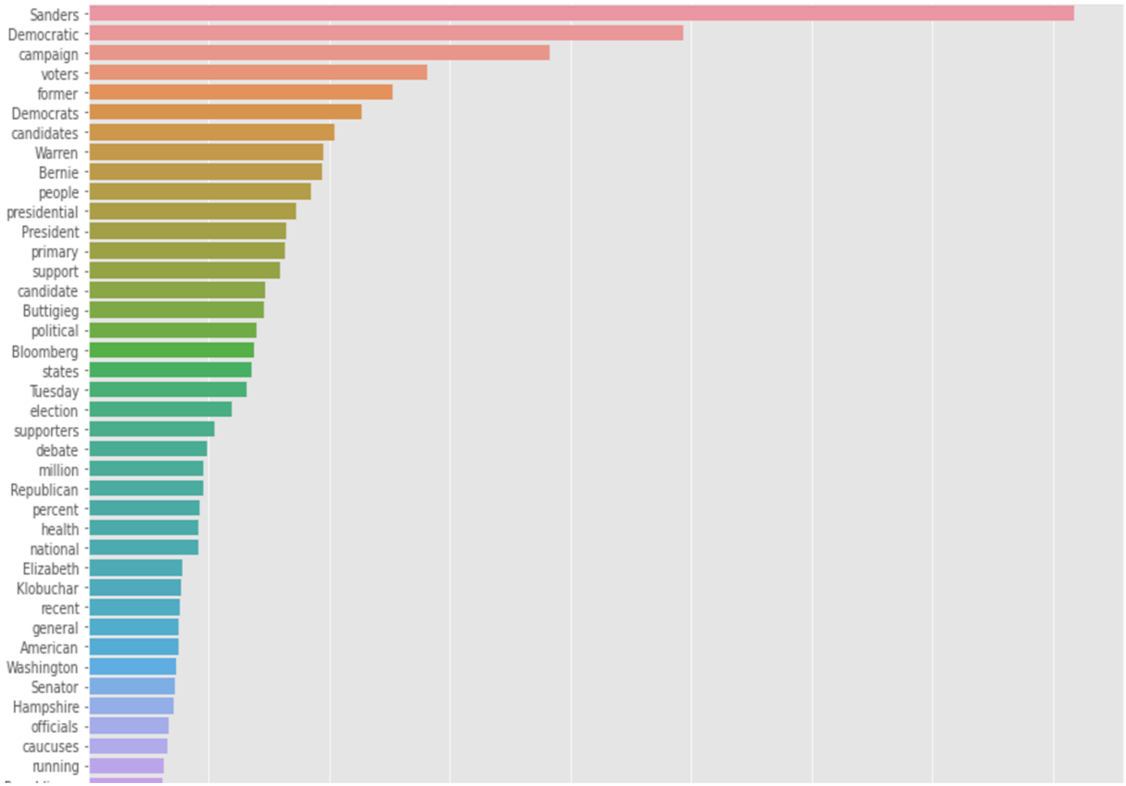


MOST FREQUENT WORDS ABOUT EACH CANDIDATE
(TOP 50 WORDS, WORD LENGTH>5)

DONALD TRUMP


campaign
Democratic
voters
former
Democrats
President
presidential
people
political
candidates
Warren
primary
states
election
support
candidate
Tuesday
Senate
impeachment
Republican
Bernie
Bloomberg
Buttigieg
supporters
national
debate
Republicans
million
health
Senator
officials
American
voting
Washington
recent
Hampshire
policy
general
public
percent

JOE BIDEN



BERNIE SANDERS

Information Extraction with Textacy

Here are the things we learned about **Trump**:

- a fighter
- a Racist President
- reluctant to even hear about election interference, and Republicans dislike discussing it publicly
- not the right person for the job
- almost uniformly in agreement that he should be removed for his behavior
- not wrong
- racist and that he had helped to make racism a bigger problem in the country
- on uncertain political ground
- the first impeached president ever to seek re-election

Here are the things we learned about **Biden**:

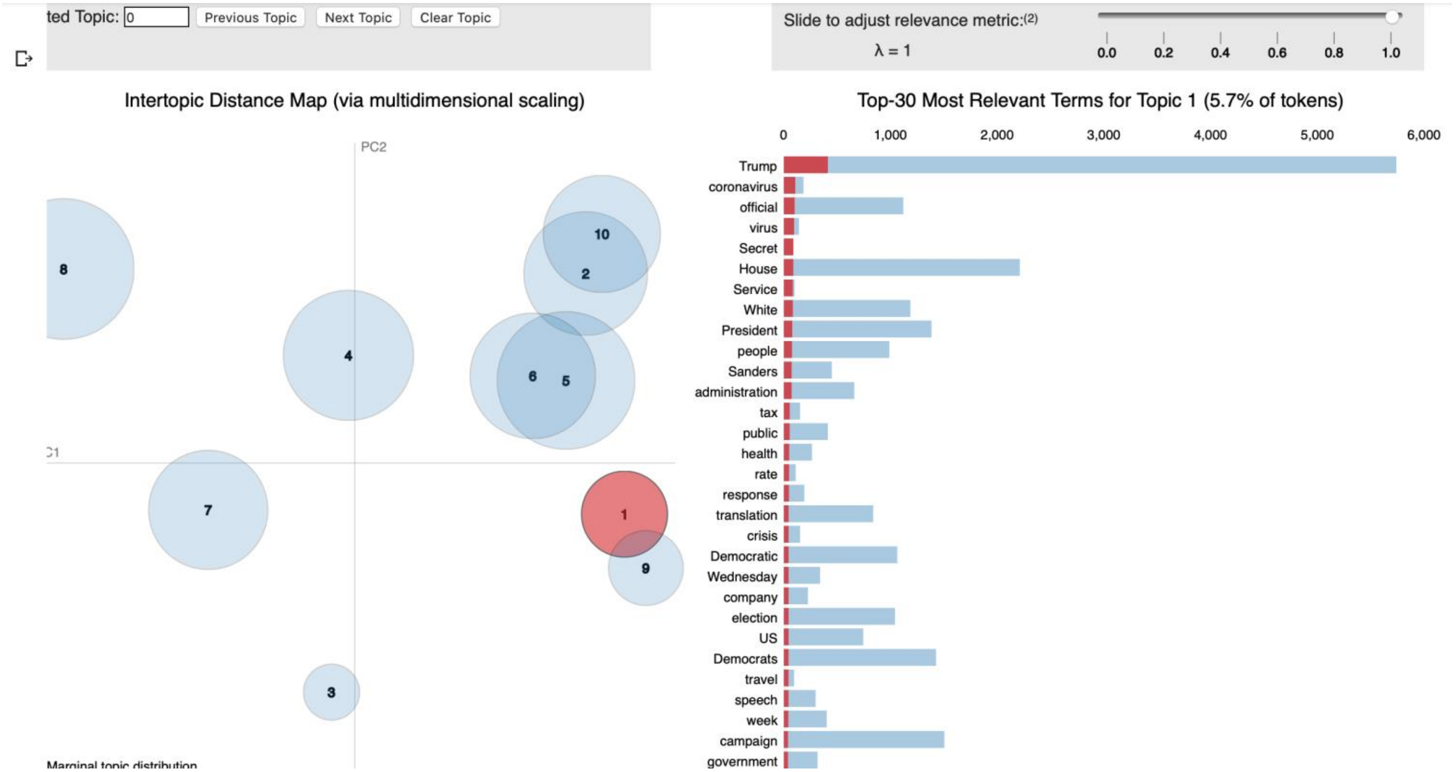
- among its most vocal Democratic supporters
- not shy about defending other elements of his record from his liberal rival
- right to be more assertive in contrasting himself with Mr. Buttigieg and Mr. Sanders
- the man to meet the moment
- likely to build an insurmountable delegate lead over the next few weeks
- the viable moderate candidate
- the party's best nominee
- the candidate who should win

Here are the things we learned about **Sanders**:

- particularly direct with his criticism, and he and Ms. Warren have accused Mr. Bloomberg of trying to buy the election
- in a "uniquely bad position" to win a general election
- likely to have the money to keep running if he wants to
- still in the presidential race - what they signed up for
- the 2020 embodiment of Dr. Martin Luther King Jr.'s political ideology
- n't a Democrat
- popular with young voters

TOPIC MODELING WITH GENSIM

(TRUMP EXAMPLE)



(0, '0.021*"Trump" + 0.006*"coronavirus" + 0.005*"official" + 0.005*"virus" + 0.005*"Secret" + 0.005*"House" + 0.004*"Service" + 0.004*"White" + 0.004*"President" + 0.004*"people"')

(1, '0.014*"Trump" + 0.013*"campaign" + 0.007*"state" + 0.006*"Democratic" + 0.006*"Iowa" + 0.006*"voter" + 0.005*"Biden" + 0.004*"President" + 0.004*"Sanders" + 0.004*"candidate"')

(2, '0.018*"Parnas" + 0.014*"Trump" + 0.007*"Giuliani" + 0.006*"Ukraine" + 0.005*"House" + 0.005*"Schiff" + 0.004*"impeachment" + 0.003*"lawyer" + 0.003*"Fruman" + 0.003*"White"')

(3, '0.016*"Trump" + 0.006*"case" + 0.005*"House" + 0.005*"Justice" + 0.005*"former" + 0.004*"Department" + 0.004*"Senate" + 0.004*"Stone" + 0.004*"attorney" + 0.004*"people"')

(4, '0.023*"Trump" + 0.005*"US" + 0.004*"American" + 0.004*"President" + 0.004*"people" + 0.003*"Iran" + 0.003*"year" + 0.003 *"House" + 0.003*"campaign" + 0.003*"official"')

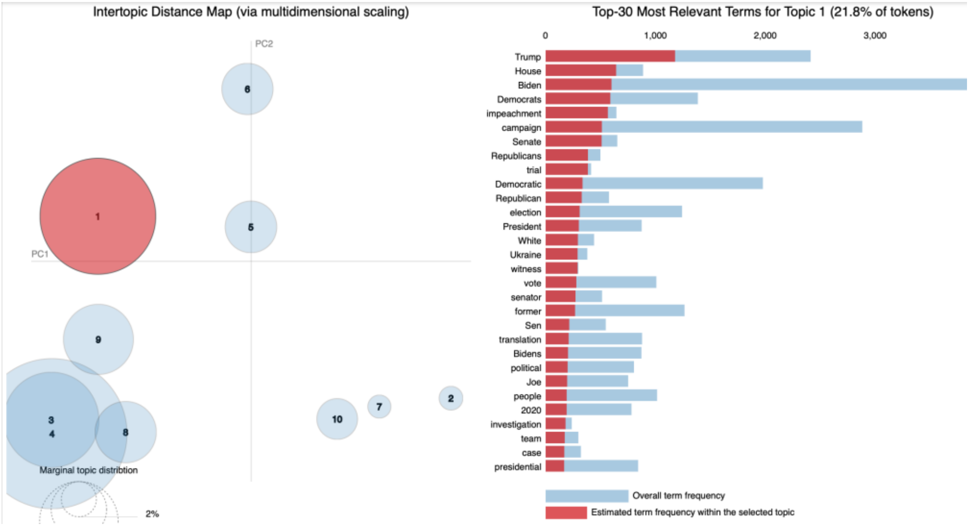
(5, '0.015*"Trump" + 0.007*"official" + 0.005*"House" + 0.005*"campaign" + 0.005*"intelligence" + 0.005*"election" + 0.004*"party" + 0.004*"Republican" + 0.004*"President" + 0.004*"administration"')

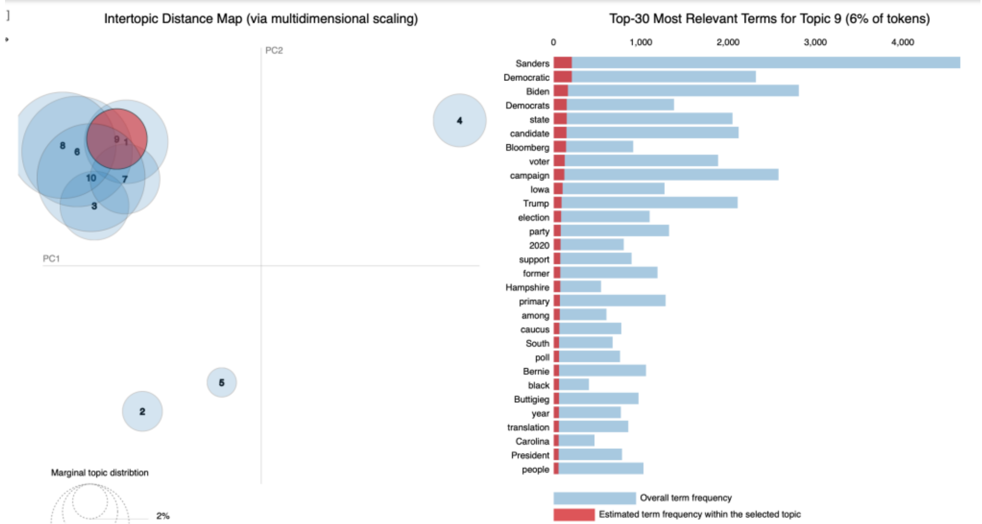
(6, '0.016*"Trump" + 0.016*"House" + 0.010*"Ukraine" + 0.010*"White" + 0.007*"impeachment" + 0.007*"official" + 0.006*"Giuliani" + 0.005*"lawyer" + 0.005*"President" + 0.004*"investigation"')

(7, '0.017*"impeachment" + 0.016*"Senate" + 0.014*"House" + 0.013*"trial" + 0.013*"Trump" + 0.012*"Democrats" + 0.009*"witness" + 0.009*"Republicans" + 0.009*"vote" + 0.008*"senator"')

(8, '0.013*"Trump" + 0.004*"market" + 0.004*"Pence" + 0.003*"President" + 0.003*"stock" + 0.003*"year" + 0.003*"Washington" + 0.002*"2020" + 0.002*"economy" + 0.002*"translation"')

(9, '0.018*"Trump" + 0.009*"campaign" + 0.007*"voter" + 0.007*"Biden" + 0.007*"state" + 0.007*"Democratic" + 0.006*"Bloomberg" + 0.005*"Democrats" + 0.005*"election" + 0.005*"candidate"')





TOPIC MODELING WITH GENSIM

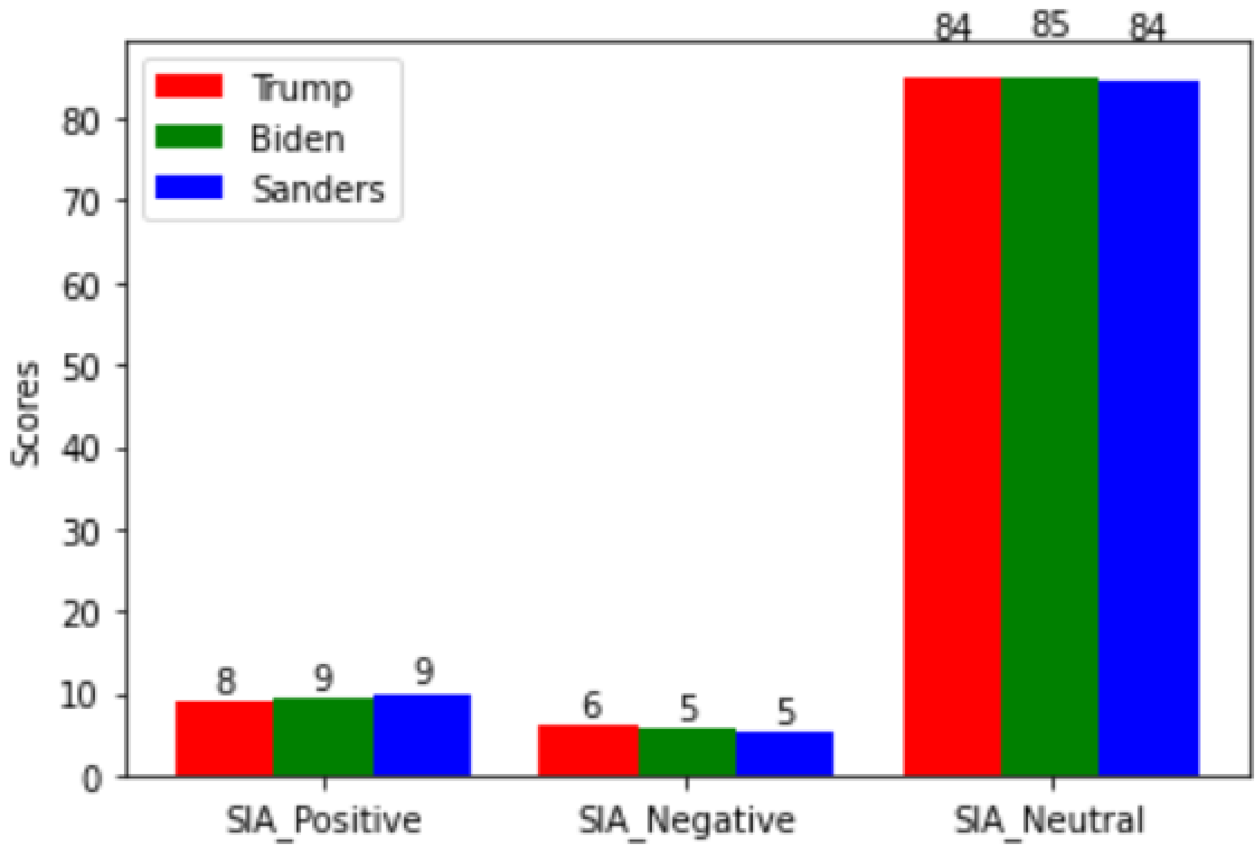
(BIDEN & SANDERS EXAMPLES)

(0, '0.016*"Trump" + 0.009*"House" + 0.008*"Biden" + 0.008*"Democrats" + 0.008*"impeachment" + 0.007*"campaign" + 0.007*"Senate" + 0.005*"Republicans" + 0.005*"trial" + 0.004*"Democratic"') (1, '0.003*"candidate" + 0.003*"Manchin" + 0.003*"costume" + 0.003*"say" + 0.003*"campaign" + 0.002*"Hampshire" + 0.002*"Democratic" + 0.002*"Clinton" + 0.002*"time" + 0.002*"state"') (2, '0.020*"Sanders" + 0.019*"Biden" + 0.010*"campaign" + 0.007*"candidate" + 0.007*"state" + 0.006*"Democratic" + 0.006*"primary" + 0.005*"debate" + 0.004*"former" + 0.004*"Warren"') (3, '0.015*"Sanders" + 0.014*"Biden" + 0.011*"voter" + 0.010*"state" + 0.008*"campaign" + 0.007*"candidate" + 0.007*"Democratic" + 0.006*"primary" + 0.005*"Tuesday" + 0.004*"vote"') (4, '0.007*"Trump" + 0.006*"Biden" + 0.005*"campaign" + 0.004*"candidate" + 0.004*"US" + 0.004*"Democratic" + 0.004*"Security" + 0.003*"official" + 0.003*"House" + 0.003*"presidential"') (5, '0.006*"Trump" + 0.006*"campaign" + 0.006*"Parnas" + 0.006*"company" + 0.005*"Biden" + 0.005*"message" + 0.005*"Ukraine" + 0.004*"election" + 0.004*"political" + 0.004*"Russian"') (6, '0.009*"campaign" + 0.006*"million" + 0.006*"candidate" + 0.004*"Bloomberg" + 0.003*"Democratic" + 0.003*"money" + 0.003*"presidential" + 0.003*"state" + 0.003*"raised" + 0.002*"Biden"') (7, '0.012*"Iowa" + 0.010*"woman" + 0.010*"campaign" + 0.008*"Biden" + 0.008*"candidate" + 0.007*"Warren" + 0.006*"Democratic" + 0.006*"caucus" + 0.004*"voter" + 0.004*"people"') (8, '0.017*"campaign" + 0.010*"Bloomberg" + 0.008*"state" + 0.008*"Trump" + 0.008*"Democratic" + 0.007*"Biden" + 0.006*"million" + 0.005*"candidate" + 0.005*"Facebook" + 0.005*"political"') (9, '0.006*"Iowa" + 0.006*"Trump" + 0.003*"union" + 0.003*"Americans" + 0.003*"state" + 0.003*"year" + 0.003*"candidate" + 0.003*"black" + 0.003*"people" + 0.003*"presidential"')

(0, '0.012*"Biden" + 0.012*"campaign" + 0.009*"Sanders" + 0.009*"Trump" + 0.007*"Democratic" + 0.006*"state" + 0.006*"candidate" + 0.006*"Democrats" + 0.004*"voter" + 0.004*"former"') (1, '0.008*"Trump" + 0.006*"Biden" + 0.005*"Sanders" + 0.004*"people" + 0.004*"Tuesday" + 0.004*"vote" + 0.003*"Sen" + 0.003*"House" + 0.003*"voted" + 0.003*"voter"') (2, '0.011*"Sanders" + 0.009*"campaign" + 0.007*"candidate" + 0.006*"Democratic" + 0.005*"million" + 0.004*"Warren" + 0.004*"Biden" + 0.004*"Iowa" + 0.004*"voter" + 0.004*"political"') (3, '0.018*"Sanders" + 0.008*"tax" + 0.007*"policy" + 0.005*"Democratic" + 0.005*"business" + 0.005*"government" + 0.005*"Sanderss" + 0.004*"market" + 0.004*"health" + 0.004*"economist"') (4, '0.008*"state" + 0.005*"Buttigieg" + 0.005*"election" + 0.004*"Democratic" + 0.004*"caucus" + 0.004*"Iowa" + 0.003*"convention" + 0.003*"voter" + 0.003*"percent" + 0.003*"group"') (5, '0.015*"Sanders" + 0.009*"voter" + 0.009*"candidate" + 0.008*"state" + 0.008*"Biden" + 0.008*"Democratic" + 0.007*"Iowa" + 0.006*"Trump" + 0.006*"Buttigieg" + 0.006*"campaign"') (6, '0.016*"Sanders" + 0.008*"campaign" + 0.006*"Biden" + 0.006*"Warren" + 0.005*"Bloomberg" + 0.005*"candidate" + 0.004*"debate" + 0.004*"Trump" + 0.004*"Democratic" + 0.004*"Bernie"') (7, '0.016*"Sanders" + 0.014*"Biden" + 0.010*"campaign" + 0.008*"state" + 0.007*"Democratic" + 0.007*"voter" + 0.007*"Tuesday" + 0.006*"Trump" + 0.006*"primary" + 0.006*"candidate"') (8, '0.010*"Sanders" + 0.010*"Democratic" + 0.008*"Biden" + 0.007*"Democrats" + 0.007*"state" + 0.007*"candidate" + 0.007*"Bloomberg" + 0.006*"voter" + 0.006*"campaign" + 0.005*"Iowa"') (9, '0.014*"Sanders" + 0.008*"Trump" + 0.008*"Democratic" + 0.007*"campaign" + 0.006*"Iowa" + 0.006*"state" + 0.006*"election" + 0.006*"candidate" + 0.005*"party" + 0.005*"caucus"')

SENTIMENT INTENSITY ANALYZER

	sia_positive	sia_negative	sia_neutral	sia_compound
Sanders	9.904222	5.417778	84.674000	0.862205
Biden	9.393333	5.545556	85.062222	0.754183
Trump	8.795111	6.300889	84.904889	0.517104



TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

	Biden Terms	Biden TF-IDF Score	Sanders Terms	Sanders TF-IDF Score	Trump Terms	Trump TF-IDF Score
1	ms	0.040405	bloomberg	0.046533	biden	0.040791
2	bloomberg	0.037313	ms	0.043925	bloomberg	0.027297
3	impeach	0.035570	percent	0.033493	investig	0.026325
4	buttigieg	0.031897	million	0.026562	ukrain	0.025381
5	south	0.027681	super	0.025339	ms	0.024838
6	carolina	0.027412	nevada	0.024678	case	0.023303
7	black	0.025652	health	0.024547	voter	0.023002
8	trial	0.024963	deleg	0.024390	iowa	0.022930
9	super	0.024686	carolina	0.024179	wit	0.022819
10	ad	0.022953	klobuchar	0.023548	bolton	0.022006
11	million	0.022014	debat	0.023396	sander	0.020521
12	deleg	0.021646	black	0.020558	parti	0.020516
13	percent	0.021538	caucus	0.020429	team	0.020282
14	health	0.021326	hous	0.019693	sen	0.019867
15	debat	0.020395	night	0.019591	schiff	0.019792
16	michigan	0.020139	result	0.018803	tuesday	0.019167
17	klobuchar	0.019518	clinton	0.018635	ad	0.018185
18	offici	0.019255	ad	0.018307	giuliani	0.018032
19	party	0.019025	rais	0.018281	romney	0.018000
20	obama	0.018271	hampshir	0.018277	remov	0.017443

Model Architecture

Prediction Goal: predicting the sentiments of each news article

Samples: 60 articles for each candidate; 180 hand annotated articles in total

Target Labels: negative (0), neutral (1), positive (2); balanced dataset

Classifiers: XGBoost

Parameters	DTM Variations		
Term Weights	Binary Count Occurrence	TF	TF-IDF
N-Gram	Unigram	Bigram	Trigram

Model Evaluation

	config	accuracy	f1-score
0	unigram, binary	0.4167	0.4091
1	bigram, binary	0.4167	0.4139
2	trigram, binary	0.4167	0.3969
3	unigram, TF	0.3611	0.3563
4	bigram, TF	0.3611	0.3484
5	trigram, TF	0.3889	0.3790
6	unigram, TF-IDF	0.4444	0.4437
7	bigram, TF-IDF	0.4167	0.4091
8	trigram, TF-IDF	0.4167	0.4051

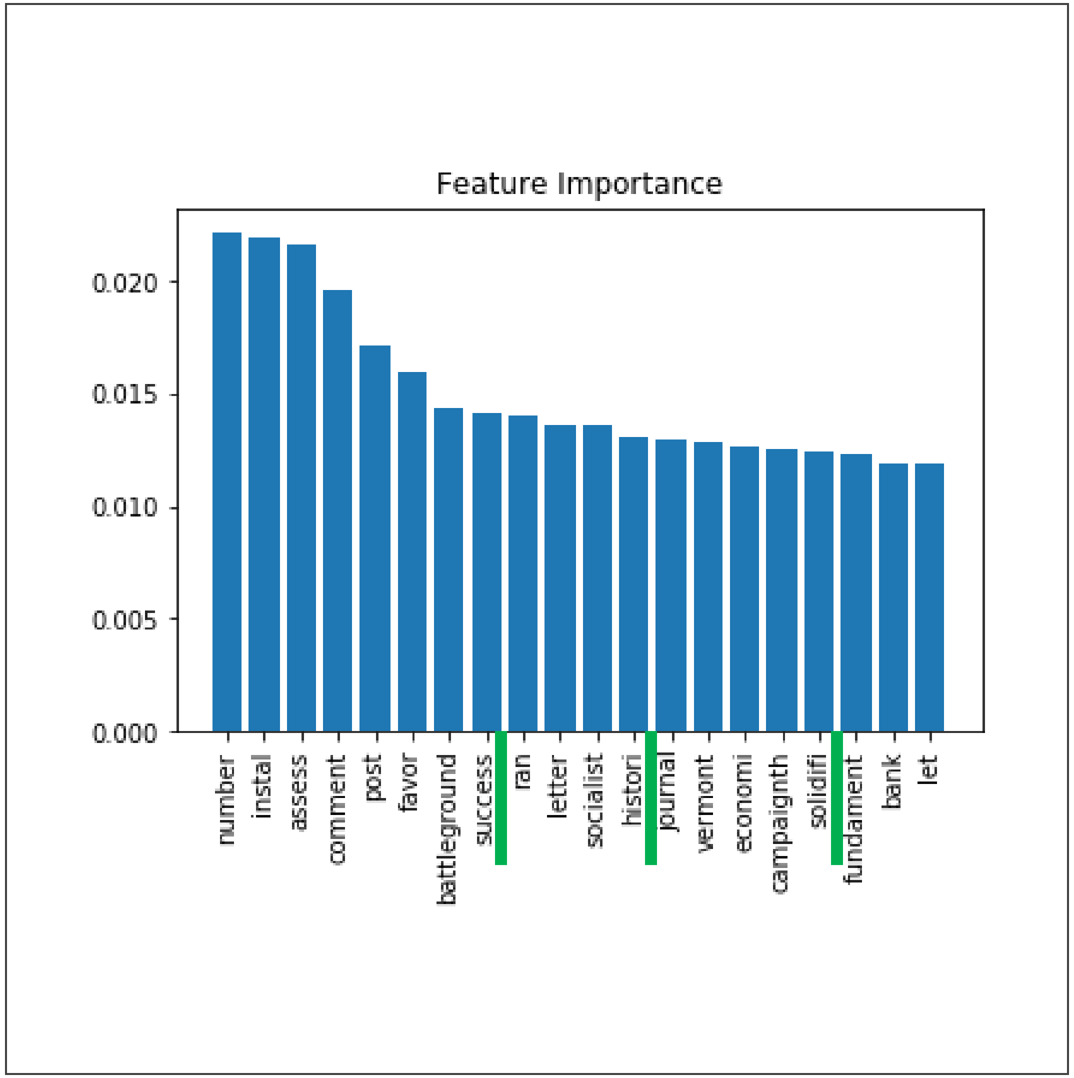
Figure 1. Training Results

Macro f1-score: 0.4437

	precision	recall	f1-score	support
0	0.38	0.42	0.40	12
1	0.58	0.58	0.58	12
2	0.36	0.33	0.35	12
accuracy			0.44	36
macro avg	0.44	0.44	0.44	36
weighted avg	0.44	0.44	0.44	36

Figure 2. Classification Report of the Best Model

- We selected the best model based on macro f1-score (unigram, TF-IDF)
- The best model is more successful in identifying neutral sentiment



Model Evaluation

- Feature importance revealed notable words driving the classification mechanism
  - I.e. success, histori, solidifi
- Model Limitations:
  - Small training set
  - Annotation criteria is subjective and inconsistent
  - DTMs did not highlight sentiment-related words enough

	Joe Biden	Bernie Sanders	Donald Trump
TF	Hard to tell the differences (basic identifiers)		
TF - IDF	Hard to tell the differences (key political events)		
Textacy Statement Extraction	Mostly Positive	Neutral	Mostly Negative
Gemsim Topic Modeling	Slightly Negative	Neutral	Slightly Negative
Sentiment Intensity Analyzer	Positive	Positive	Positive

Table. 1 Summary of Different NLP Methods

Conclusion

- Sentiment differences using different NLP information extraction methods (Table. 1)
- Trump is mostly negatively described in news articles; Sanders maintained a neutral reputation; Biden had a mix of sentiments.
- Sentiment Classification Model: Based on f1-score, the best model configuration is to use unigram tokens and TF-IDF scores to convert document into numeric representation.
- Model Limitation: sample size, manual annotation method, word embedding

References

Blanco, José "Hacking Scikit-Learn's Vectorizers." *Medium*, Towards Data Science, 15 Feb. 2018, towardsdatascience.com/hacking-scikit-learns-vectorizers-9ef26a7170af.

Brownlee, Jason. "How to Prepare Text Data for Machine Learning with Scikit-Learn." *Machine Learning Mastery*, 7 Aug. 2019, machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/.

Dwivedi, P. (2018). *NLP: Extracting the main topics from your dataset using LDA in minutes*. Retrieved from https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

Maklin, Cory. "TF IDF: TFIDF Python Example." *Medium*, Towards Data Science, 21 July 2019, towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76.

Malik, U. (n.d.). *Python for NLP: Sentiment Analysis with Scikit-Learn*. Retrieved from https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/

Nag, Avishek. "Text Classification by XGBoost & Others: A Case Study Using BBC News Articles." *Medium*, Towards AI-Multidisciplinary Science Journal, 3 July 2019, medium.com/towards-artificial-intelligence/text-classification-by-xgboost-others-a-case-study-using-bbc-news-articles-5d88e94a9f8.

Singh, Mahima. "Trump and the Media." *News Sentiment Analysis*, amiham-singh.github.io/.

Project Code: https://github.com/weining20000/NLP_The-2020-Presidential-Race

Thank you!