Understanding Public Opinion Through White House Press Briefings:

An Application of Natural Language Processing and Unsupervised Clustering Learning Methods

Wing Shan Chan

Northwestern University School of Professional Studies

Master of Science in Data Science Program

https://www.linkedin.com/in/wingschan/

June 1, 2022

Abstract

Public opinion is the foundation of our democracy, and it is of interest to many in the political world to better understand it. In this study, we described an explorative effort to gauge the evolving public opinion in the United States via a data science-based examination of White House press briefings under the Biden Administration.

A corpus of question and answer (Q&A) exchanges between Press Secretary Ms. Jen Psaki and the press over the course of 14 months was curated. We processed and vectorized the corpus using Natural Language Processing techniques such as spaCy part-of-speech tagging and doc2vec. Key topics of public interest were extracted to gain insights on the driving force behind presidential job approval rating movements. We investigated unsupervised clustering methods including k-means clustering and spectral biclustering in search for an optimal way to cluster the corpus. Consensus scores based on the biclusters generated between the Q&A were used to gauge coherence between the press and Ms. Psaki. Sentiments from the Q&A exchanges were analyzed. We found the topics on which the press and Ms. Psaki shared common sentiments to be those earning public approvals.

This study aimed to establish a foundation for an objective and scientific way to interpret public opinion, which is much needed in this age of misinformation.

Keywords: Natural language processing, Doc2Vec, text vectorization, spaCy, lemmatization, part of speech tagging, unsupervised machine learning, k-means clustering, spectral biclustering, sentiment analysis, VADER, White House, press briefing, Biden, Psaki, presidential job approval rating, politics, public opinion.

1. Introduction

Public opinion is crucial to democracy in the United States (Mohamed 2021). For a publicly elected government in a country with freedom of speech, public opinion represents the minds of the people. It signifies the general sentiment of the society towards the government and towards the world we live in. It guides the policymaking process and sets governmental goals (Stimson 1999). It sways election results for politicians.

Many existing endeavors aim to understand public opinion. There are forums in public media that focus on dissecting policies. There are public polls on topics from economic outlook to voting preferences. There are election outcomes which are voices from the voters. With fewer people willing to answer polls (Byler 2021) and widespread misinformation in social media (Bordalo, Yang, and Doepke 2021), getting an accurate grasp of public opinion seems increasingly difficult.

We propose a new method to interpret public opinion in the United States. White House Press Secretary Ms. Jen Psaki has been holding press briefings almost every weekday where she answered questions from the press. The press briefings serve as a way of communication between the Biden Administration and the general public. Unlike a presidential address or an op-ed in a newspaper, this channel of communication is two-way. While the press briefings are primarily for the Administration to respond to questions from the press, the questions being asked are often telling of the political climate and public interest at the time.

In this study we aim to formulate a dataset from White House press briefing transcripts of the Biden Administration between its inauguration in January 2021 to March 2022, to interpret public sentiment from the data by deploying natural language processing (NLP) techniques such as doc2vec, and to obtain insights in relations to presidential job approval ratings through unsupervised clustering techniques.

2. Literature Review

President Joe Biden, as the leader of the United States, is the most visible person representing his Administration. While the governing is done by many offices and branches of the government, President Biden's approval rating is indicative of how the US people feel about the state of affairs and their optimism towards the future. Presidential job approval ratings are primarily measured and tracked by polls conducted by various organizations such as Reuters (Reuters 2022), The Economist/YouGov (YouGov, n.d.) and Rasmussen Reports (Rasmussen Reports, n.d.), and they all track metrics of both approval and disapproval ratings. While these polls are conducted through various means such as automated or operator assisted phone polling and online polling, and may hit a different sample of demographics in terms of age, gender, race, and politics, the general trend of the ratings appear largely similar. FiveThrityEight consolidates the major polls on the matter with a weighted approach favoring high quality polls, and arrives at an averaged approval and disapproval rating trend for President Biden (Rakich 2021) as seen in Figure 1.

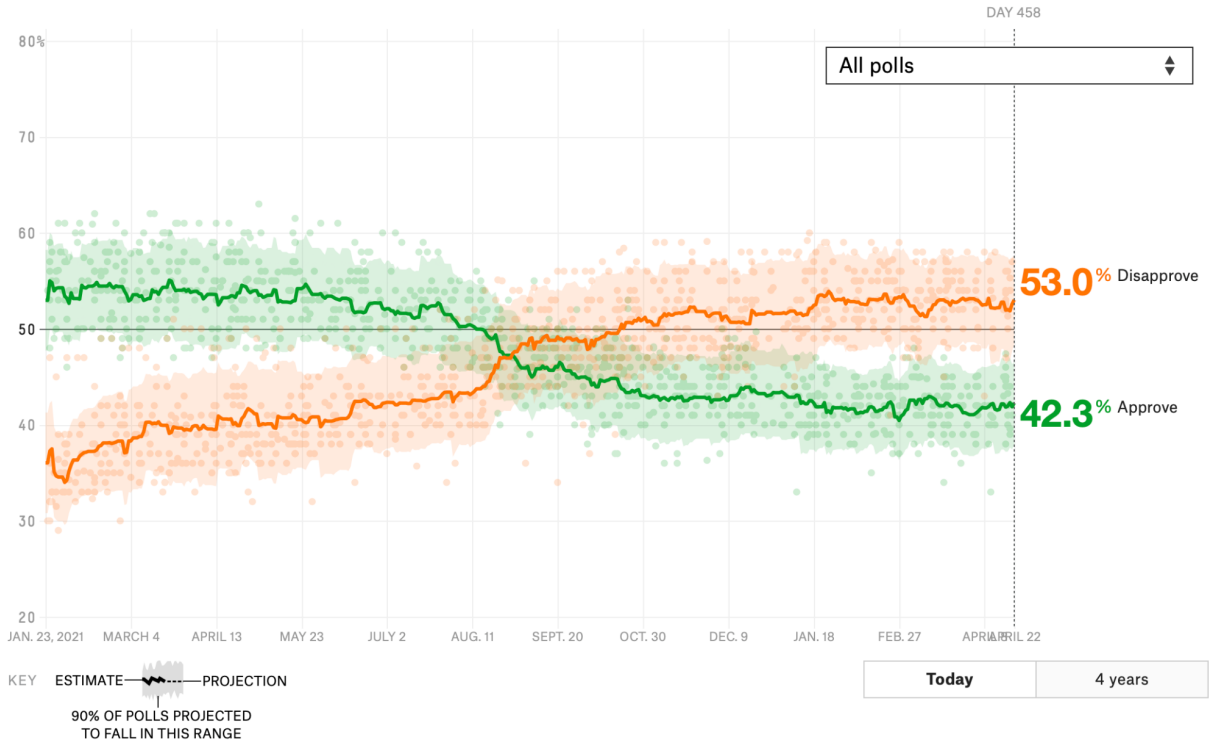


Figure 1 *Presidential job approval and disapproval ratings of President Joe Biden.* Data from January 2021 to April 2022 obtained and weighted by FiveThirtyEight (FiveThirtyEight 2022).

Most of the data-related effort surrounding the interactions between White House and the press involves archiving past briefing transcripts and counting the number of briefings conducted. The National Archive has archived past Presidential White House websites (National Archive, n.d.). Some past administrations also keep an online record of their press briefings (Bush White House Archives, n.d.; Obama White House Archives, n.d.). The frequencies of direct press briefings between Presidents and the press were recorded by presidencies and by calendar year (UC Santa Barbara, n.d.). For analysis on the briefings, news media summarize and comment on press briefings on their editorial page (PBS, n.d.; Politico, n.d.). For language analysis, a manual effort was done to analyze the words used by former President Trump during the early stage of the Coronavirus pandemic (Haberman, Peters, and Plott 2020). NLP

approaches such as term-frequency inverse-document-frequency (TF-IDF) were deployed to extract top keywords from Trump's remarks and Valence Aware Dictionary and sEntiment Reasoner (VADER) was used to gauge sentiments (jjgoings 2020). A separate effort was made in comparing the content from the Trump Administration White House website and the Obama Administration White House website via web scraping (Shaffer 2017).

In the arena of NLP to analyze the meaning of human languages, word2vec and doc2vec are popular methods to employ. Word2vec is an unsupervised learning method to represent words in a multidimensional vector space where similar words would be close to each other (Mikolov et al. 2013). Doc2vec is built upon word2vec to represent sentences and documents in a similar manner (Le and Mikolov 2014). VADER is another tool to analyze sentiment of texts and to score how positive or negative they are (Hutto and Gilbert 2014). To support these learning methods, other tools such as spaCy can be used to categorize texts into parts of speech (spaCy, n.d.).

In an effort to categorize unlabeled texts based on similarity, unsupervised machine learning methods such as clustering are often used. K-means clustering is a popular method to quantify and group data together (MacQueen 1967, 281-297), and it can be used in combination with text vectorizations to cluster texts. Spectral clustering is another popular method to reduce dimensionality and cluster data, and it is especially useful when the data are non-convex (Ng, Jordan, and Weiss 2001, 849–856; Scikit-learn, n.d.). Spectral biclustering is a method built on that. It simultaneously performs clustering on both the rows and columns of a data matrix under the presumption that the underlying structure is one of a checkerboard (Kluger et al. 2003, 703-716).

3. Data

The text transcripts of all White House press briefings from the Biden Administration are available on the White House website (The White House, n.d.), however they are in page formats by each briefing. There are no existing datasets of press briefing transcripts for data science purposes. In alignment with our objective of laying the foundation for scientific extraction of public opinion, we curated a dataset of Q&A pairs which is available upon request to the author. Table 1 summarized the curated dataset. Below we described the process in which the dataset was curated.

White House Press Briefing Q&A Dataset	
Description	Q&A exchange pairs between the press and Press Secretary Ms. Jen Psaki
Format	JSON
Timeframe	January 20, 2021 - April 7, 2022
Number of data elements	11669
Number of press briefings	456

Table 1 *Summary of curated dataset of White House press briefing Q&A pairs.* The dataset is available upon request to the author.

3.1 Web Scraping

We employed web scraping techniques to extract the transcripts into usable form for NLP processing. To do this we first examined what is available on the White House website. On the website, the Administration provides blogs, legislation, speeches, statements, and press briefing transcripts. Since we are interested in the interactive nature of the press briefings, we focused our web scraping on that only.

The Press Briefings pages each have a maximum of 10 briefings listed with clickable links into each of them, as seen in Appendix 1. The Administration has over 400 briefing transcripts as of the time this report is written. The URL for each of these pages was found to be https://www.whitehouse.gov/briefing-room/press-briefings/**1,** with the last digit incrementing for the next 10 listings. When the listings are exhausted, the website would return a "404 not found" message.

In order to get the content of the briefings, we first used Python libraries Beautiful Soup and Requests to collect the URL for each of the briefings by crawling through the listings until we received a "404 not found" message. Each URL is retrieved by getting the href text from the uniquely identifying class of the briefing title. After getting the list of URLs for the briefings, we reversed the list so that we have a chronological order of briefings. We requested them one by one and extracted the useful information that we wanted. We used the order as "id" for the document, and we had the URL as an item. We collected the title of the briefing, the date of the briefing, and the body of the briefing.

3.2 Question and Answer (Q&A) Exchange Extractions

The body of the briefing details all that was said during the briefing, which was further processed to extract the Q&A exchange between Press Secretary Ms. Psaki and the press. An example is listed in Appendix 2. In a typical briefing, Ms. Psaki usually opens with an initial statement and we excluded it since it is not an exchange with the press. The briefing would then proceed with the Q&A exchange, which were always preceded by "Q:" and "MS PSAKI:" in the transcript. An example is shown in Appendix 3. The Q&A pairs were extracted with the help of

regular expressions after basic text cleaning to remove any punctuation, line breaks or Unicode. The date of which the briefing happened was used to calculate how many months into the Administration was the question asked.

One of the challenges in extracting the exchange between Ms. Psaki and the press was that sometimes Ms. Psaki would bring in a guest such as Dr. Fauci to speak on specific topics, for example, the coronavirus pandemic. Since we are aiming to gauge public opinions by examining what was being asked in general at a press briefing, any exchanges answered by people other than Ms. Psaki were discarded.

3.3 Press Briefing Dataset Results

The results were then stored into a JSON file. An example of the processed entry is shown in Figure 2. Between the inauguration on January 20 2020 and April 7 2022, the data extraction yielded 456 press briefings with 11669 exchanges between Ms. Psaki and the press. For this work, we excluded any exchanges that had fewer than five words on both sides to remove casual greetings and acknowledgements. We arrived at 8322 Q&A pairs for our study.


```
{'id': 4,
 'briefing_id': 0,
 'qna_id': 4,
 'url': 'https://www.whitehouse.gov/briefing-room/press-briefings/2021/01/20/press-briefing-by-
press-secretary-jen-psaki-january-20-2021/',
 'title': 'Press Briefing by Press Secretary Jen Psaki, January 20, 2021',
 'date': 'January 20, 2021',
 'DayInOffice': 0,
 'MonthInOffice': 0,
 'press': 'with regards to reopening schools what level of vaccination in teachers or students
or level of testing does the administration think would be appropriate in order to meet the
target date that the president has said',
 'psaki': 'this is a great question and as i noted at the beginning as a mom myself i want to
know all the details as well were going to have more to share from our health experts in the
coming days and i will venture to get them in here to give you all a briefing on the specifics
but we really want to lean into them on their expertise on that front go ahead ill come right to
you right next go ahead'}
```


Figure 2 *Example entry of the extracted corpus.* An entry was obtained after data processing to extract Q&A exchanges between Ms. Psaki and the Press.

3.4 Movement categorization of Presidential Job Approval Ratings

Part of this study involved investigating any relationship between the extracted sentiments from the press briefing transcripts and the weighted presidential job approval ratings from FiveThirtyEight (FiveThirtyEight 2022), which serves as an indicator for public opinion on the state of affairs. Considering the variation in daily number of exchanges we have in our press briefing dataset due to our discarding of exchanges not involving Ms. Psaki, we opted to aggregate the ratings into weekly averages and to classify the weekly change as categories as followed.

Increase: (Current week average - last week average) >= 0.25%

Neutral: (Current week average - last week average) between -0.25% and 0.25%

Decrease: (Current week average - last week average) <= -0.25%

The categories were calculated using the data from FiveThirtyEight, and integrated into our dataset. Figure 3 showed the distribution of weekly rating changes out of the 64 weeks of data collected. Table 2 showed the count in each category.

	Approval rating weekly average change	Disapproval rating weekly average change
Increase	17	29
Neutral	17	23
Decrease	30	12

Table 2 *Number of weeks with increase, neutral and decrease in approval and disapproval rating changes*.

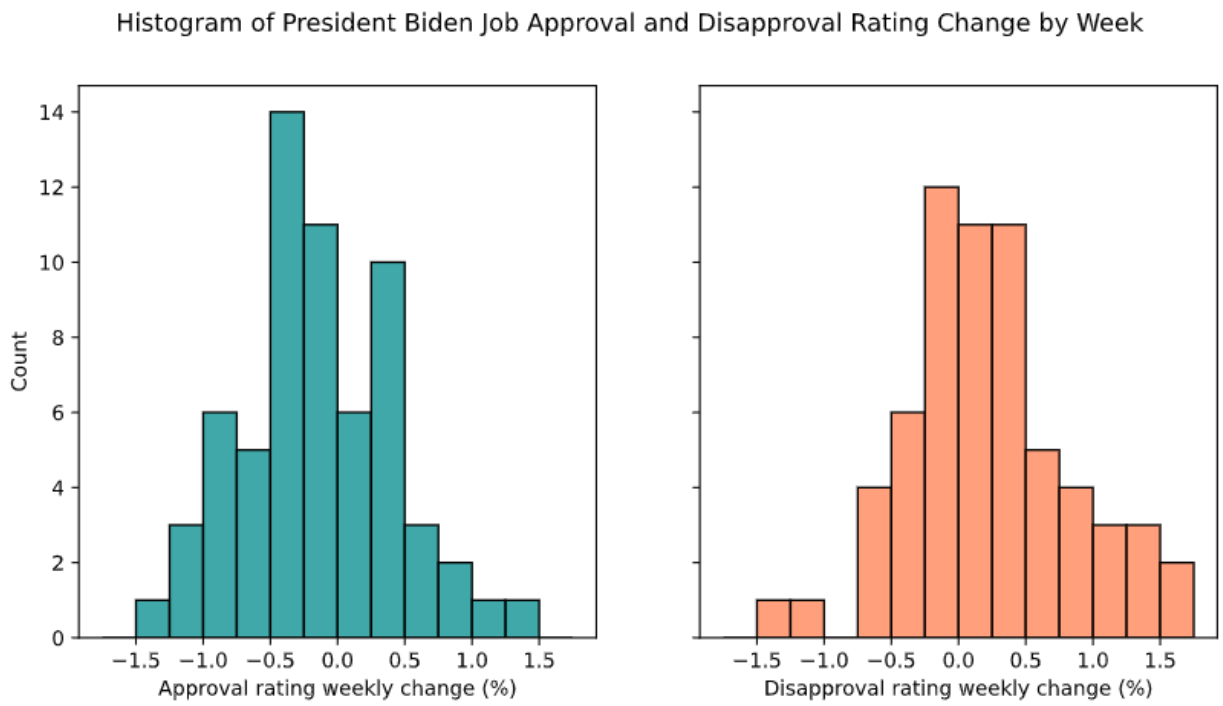


Figure 3 *Histogram of President Biden's job approval and disapproval weekly rating change*. A total of 64 weeks of data were collected.

4. Methods

Our strategy in building a model from the press briefings was to vectorize the Q&A from the dataset, and to use the resulting vectors for unsupervised clustering to obtain insights.

4.1 Lemmatization, Part of Speech (POS), and Out-of-Vocabulary (OOV) Words Extraction with spaCy

Before vectorizing, we chose to create a modified set of data using the spaCy library to remove stop words and to extract lemmas and out-of-vocabulary (OOV) words. Stop words such as "a", "the", "you", do not usually add to the meaning of a text. In addition to the common stop words identified by spaCy, we also removed the following custom set of stop words relevant to

our dataset. These words were commonly used throughout the press briefings with not much contribution to the topic identification process.

Custom stop words:
['jen','thank','thanks','president','biden','white','house','administration','united','states']

Lemmas, which refer to the roots of words, help reduce variations in our corpus and generalize the data for processing purposes. For example, the word “making” can be reduced to the root “make”. OOV words are words that are not within the spaCy pipeline package. In this study, we used the package en_core_web_lg which includes 685,000 vocabulary (spaCy, n.d.). For a word to be OOV, it will likely have to be a non-standard word such as “covid”. With the lemmatization, part of speech (POS) and OOV tagging features from spaCy, we reviewed the top words by quarter in press briefings under the Biden Administration as a guidance of top public concerns over time. We examined if certain POS and OOV would be more suitable for the task of topic extraction.

4.2 Vectorization of Q&A Data

We treated each of the Q&A pairs as a document, and vectorized them into 50, 100, 200, and 300-dimensional vectors using GenSim Doc2Vec. The process was done through tagging each document with a tag (the document id), building a vocabulary through the tagged data, and training to vectorize. With this we obtained the matrices as listed in Table 3.

M1-M3 were used for topic k-means clustering. The decision to vectorize Q&A as a joined text was based on the premise that richer texts could provide more features that would aid in clustering. M4 was built for spectra biclustering, where we matched the biclusters from the press questions and Ms. Psaki’s answers.

	Description	Number of vectors	Dimensionality
M1	Vectors of original joined Q&A	8322	[50,100,200,300]
M2	Vectors of joined Q&A with stop word removal and lemmatization	8307	[50,100,200,300]
M3	Vectors of joined Q&A with proper nouns only	3112	[50,100,200,300]
M4	Vectors of original questions and answers as separate entries under a single vectorization model	16644	[300]

Table 3 *List of matrices obtained by vectorizing the Q&A dataset and its modified sets using POS/OOV extraction.*

4.3 K-means Clustering of Vectors

K-means clustering was used to cluster the three matrices (M1-M3). The clusters were visualized using t-distributed stochastic neighbor embedding (t-SNE) multidimensional scaling. We leveraged silhouette scores and silhouette plots to examine if any of the matrices would produce more clearly defined clusters than others.

4.4 Spectral Biclustering of Vectors

We applied spectral biclustering to M4, which contains vectorized press questions and Ms. Psaki's answers. We worked under the premise that Ms. Psaki's answer for each press question would be relevant, and that if we bicluster questions and answers separately and optimally we would find the bicluster sets to be matching. We checked on this premise using consensus scores with different numbers of biclusters.

4.5 VADER sentiment analysis

We applied VADER sentiment analysis to collect sentiment scores on both questions and answers from the press briefings. VADER produces positive, negative, and a neutral score on a text based on a lexicon, and the three scores together produce a compound sentiment score, which is commonly used as a benchmark for overall sentiment (Hutto and Gilbert 2014). The compound sentiment score differences were calculated for each Q&A pair, and we learned if the press and Ms. Psaki shared closer sentiments on some topics than others. Using the aggregated weekly sentiment scores, we compared with the presidential job ratings time trend.

5. Results

In this section we summarized the key results from our study of the dataset, which included top public interest extraction over the timeline of the Biden Administration, clustering results through k-means and spectral biclustering methods, and sentiment analysis using VADER.

5.1 Top Public Interest Extraction

Over the 14 months of press briefing data reviewed in this study, many major events have happened in the US and in the world. Some of these events have spurred heated debates, others have sparked fleeting public interest. It is of political interest to understand what holds public attention and causes significant public concerns.

Using spaCy we extracted sets of different POS and OOV from our dataset. We visualized the most common words in the data using word clouds. Figure 4 showed the six sets of word clouds using the original set of press questions and their modified sets of various POS and OOV.

Figure 4 *Word clouds of the set of press questions and the modified sets using POS and OOV extractions.* With POS and OOV extraction, the top words from the dataset were more indicative of the topic of discussion. Out of the six sets, proper nouns with OOV words showed the richest information for topic identification.

The original set and lemma set failed to indicate what were the topics being discussed. With POS and OOV extractions, we as human analysts were able to see clearly the topics of interest. Out of the sets being reviewed, proper nouns with OOV words showed the richest information for topic identification.

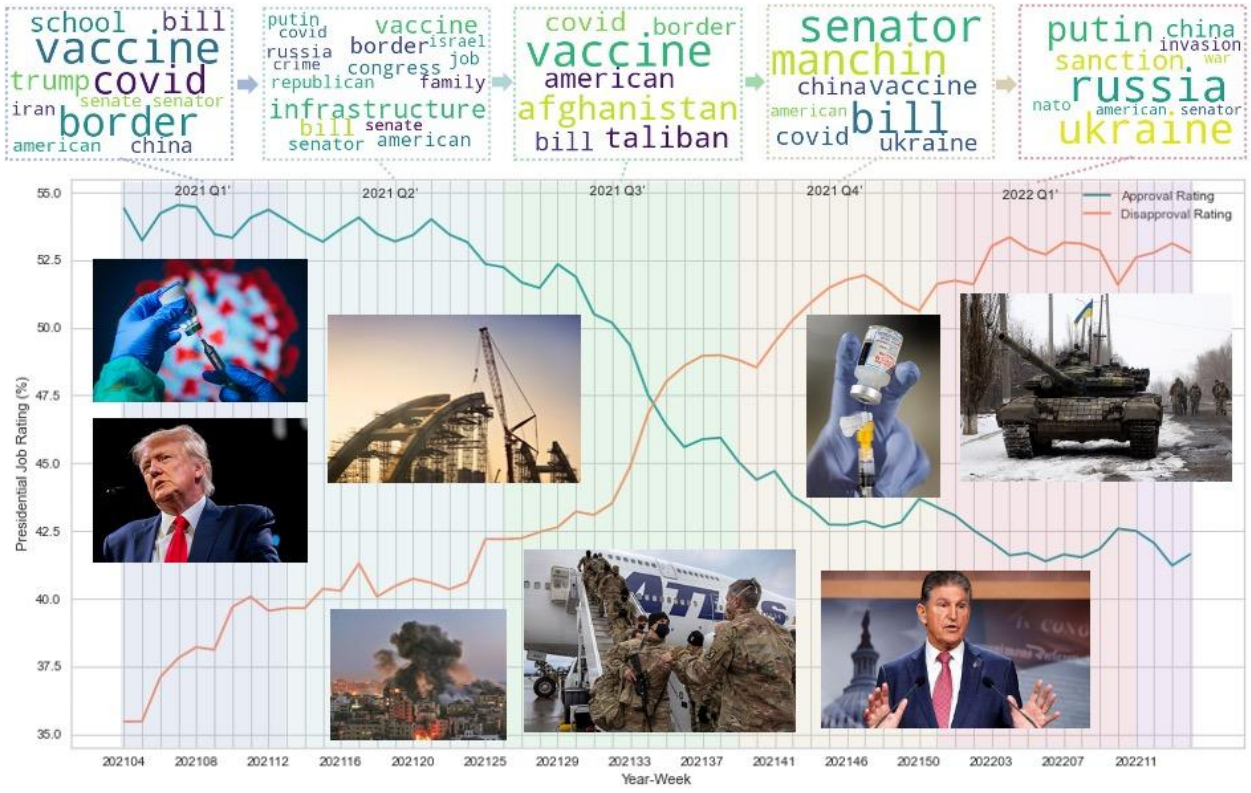


Figure 5 *Infographic showing quarterly word clouds from proper nouns with OOV words generated from press questions overlaying the presidential job approval/disapproval rating weekly averages.*

With this information we examined the top proper nouns with OOV words by quarter over the Biden Administration in an effort to establish a time trend of shifting public interests. Figure 5 showed the results of this effort overlaying a time trend of the weekly average presidential job approval and disapproval rating. Through the top proper nouns with OOV words extraction, we saw public focus shifted from covid vaccines and President Trump in Q1' 2021, to

the infrastructure bill in Q2', to US withdrawal from Afghanistan in Q3', to Senator Manchin in Q4', and to the war in Ukraine and Russia in Q1' 2022.

Equipped with the ability to extract topics through top words, we were interested in examining if certain topics were related to an increase or a decrease in job approval/disapproval ratings. Using the weekly rating categorization as illustrated in Section 3.4, we arrived at a set of categorized word clouds in Figure 6, which indicated the topics of discussion when the public is more approve or disapprove of President Biden. On the topic of the war in Ukraine and Russia, President Biden's job approval rating increased while his disapproval rating decreased. The reverse is true on the topic of Afghanistan.

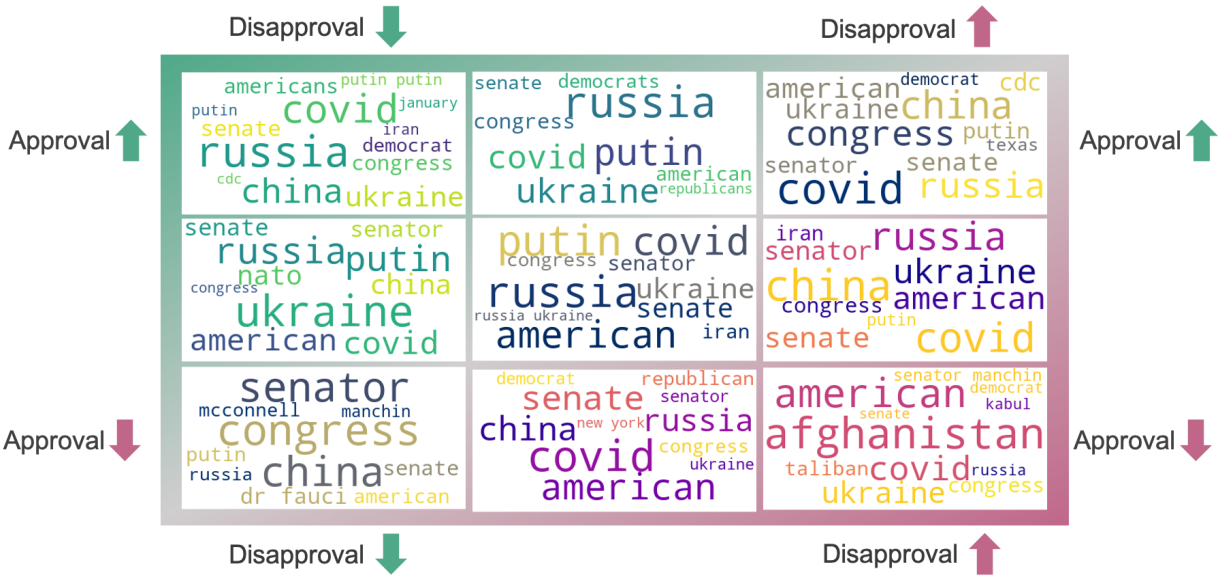


Figure 6 *Word clouds categorized by presidential job approval/disapproval rating weekly movements.* The public appeared more approve of President Biden's handling of the war in Ukraine and Russia while more disapprove of his handling of the affairs in Afghanistan.

5.2 K-means Clustering of Q&A

After vectorizing the Q&A into different sets of matrices using doc2vec as detailed in Section 4.2, we used k-means clustering method to group the Q&A together. Silhouette scores

and silhouette plots were used to judge the definition of the clusters. A higher silhouette score corresponds to more well-defined clusters. In addition, t-SNE visualizations were used to project the cluster map onto a 2-D plane.

Table 4 showed the results of the silhouette scores for M3 (proper nouns) at different dimensionality of vectorization when grouping into 5 clusters. As expected, higher dimensionality of vectorization offered more details on the texts, resulting in higher silhouette scores which indicated better clustering.

Vector dimensionality	50	100	200	300
Silhouette score	0.3515	0.3481	0.3550	0.3762

Table 4 *Silhouette scores at five clusters for M3 (proper nouns) at different vectorization dimensionality using k-means clustering.* Vectorizing at 300-dimensions showed the higher silhouette scores which indicated better clustering.

Number of clusters	M1 (original Q&A)	M2 (lemmas)	M3 (proper nouns)
3	0.165	0.1686	0.3931
4	0.1781	0.2304	0.3922
5	0.1891	0.1831	0.3762
6	0.1975	0.1918	0.38
7	0.2032	0.1982	0.3818
8	0.2079	0.2033	0.3881
9	0.2138	0.208	0.3866
10	0.2117	0.2101	0.3945
11	0.2203	0.2117	0.4041
12	0.222	0.222	0.4078

Table 5 *Silhouette scores at different numbers of clusters for M1-M3 at 300-dimensional vectors using k-means clustering.*

Table 5 showed the silhouette scores at different numbers of clusters for the matrices M1-M3 in 300-dimensional vectors, which were the original Q&A texts, lemmatized texts, and proper nouns. Silhouette scores were low for unprocessed texts in M1, slightly higher for M2 (lemmas), and with M3 (proper nouns) we obtained the highest scores.

We used the silhouette scores, the silhouette plots, and the t-SNE projections as guidance to arrive at the optimal number of clusters for each of the matrices. Figure 7 showed the silhouette plots with t-SNE projections for M1-M3. M1 was clustered into nine clusters, M2 was clustered into six, and M3 was clustered into three.

The k-means clustering results showed that M1 and M2, which had retained the most information in the vector matrices, could afford to have more number of clusters based on the level of details available. For M3 where we have extracted proper nouns only and discarded any Q&A pairs that did not contain proper nouns, the dataset size was reduced to a third. The reasonable number of clusters was smaller due to both a smaller dataset and a more limited amount of details available for clustering. It was noted that the silhouette plots were showing negative silhouette coefficients in all M1-M3, which indicated that some of the data were clustered to the wrong cluster. In all three cases, one of the clusters in the set was dominantly large with clear removal from the rest of the clusters, indicating a non-optimal clustering method based on the data. Upon inspection, there were no clear topics relating to each cluster.

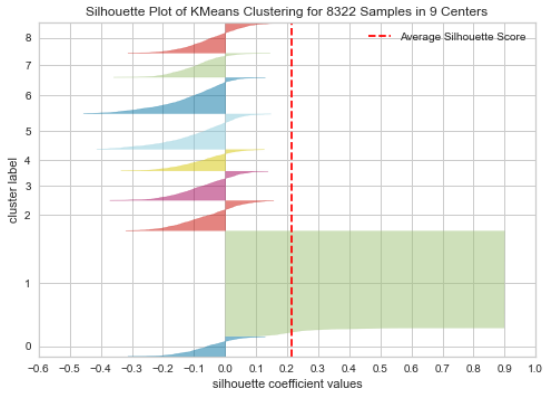
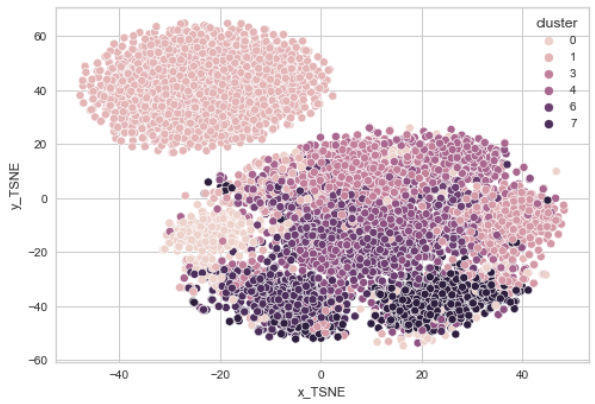
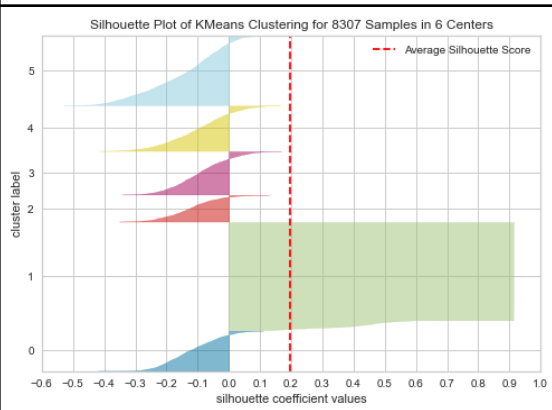
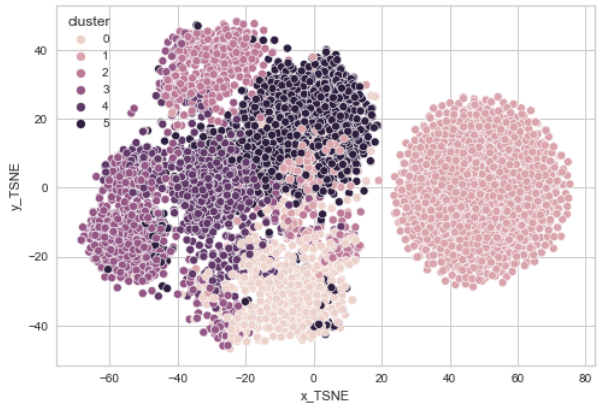
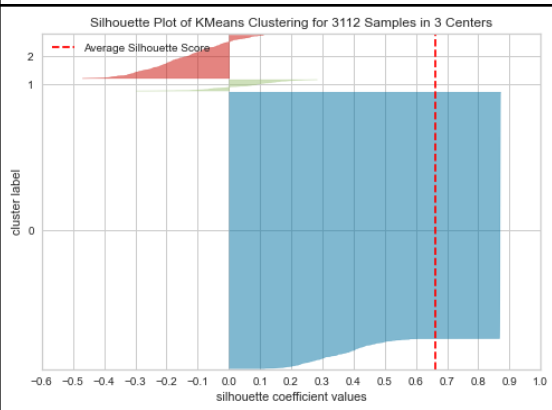
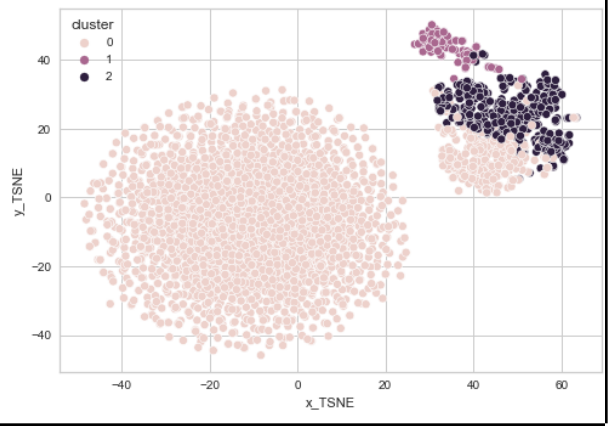
Matrix	Silhouette plot	t-SNE 2-D projection
M1 (n=9)		
M2 (n=6)		
M3 (n=3)		

Figure 7 *Silhouette plots with t-SNE projections for M1-M3 in n clusters.* M1 (original Q&A) was clustered into 9 clusters. M2 (lemmas) was clustered into 6 clusters. M3 (proper nouns) was clustered into 3 clusters.

5.3 Spectral Biclustering of Q&A

We applied spectral biclustering as a different clustering method to our data. The original questions and original answers were stacked and vectorized together at 300 dimensions to form matrix M4 using doc2vec. The questions and answers were separated after vectorization, and clustered separately using spectral biclustering. Contrasting to k-means clustering where only the rows (documents) of the matrix were clustered, spectral biclustering clustered the rows (documents) and columns (features) simultaneously, giving a bicluster structure to the data much like a checkerboard.

As our data contained a paired structure where the press and Ms. Psaki should be on the same topic in each exchange, we examined if the biclusters generated from the press questions and Ms. Psaki's answers were matching. Consensus score, which has a value between [0,1], describes the similarities between two sets of biclusters and was used to gauge if the biclusters generated between questions and answers were matching. Table 6 showed the consensus scores between the questions and answers at different numbers of row (documents) clusters and column (features) clusters. Clustering the rows into 3 groups while clustering the columns into 4 groups yielded the highest consensus score at 0.1773.

Consensus Score		Number of column (feature) clusters				
		2	3	4	5	6
Number of row (document) clusters	3	0.0832	0.1376	0.1773	0.1211	0.1359
	6	0.0423	0.0636	0.0758	0.0551	0.0552
	9	0.0287	0.0437	0.0473	0.0387	0.0390

Table 6 *Consensus scores between the press questions and Ms. Psaki's answers at different bicluster configurations using spectral biclustering.* Configuration [3,4] of 3 document clusters and 4 feature clusters yielded the best consensus score at 0.1773.

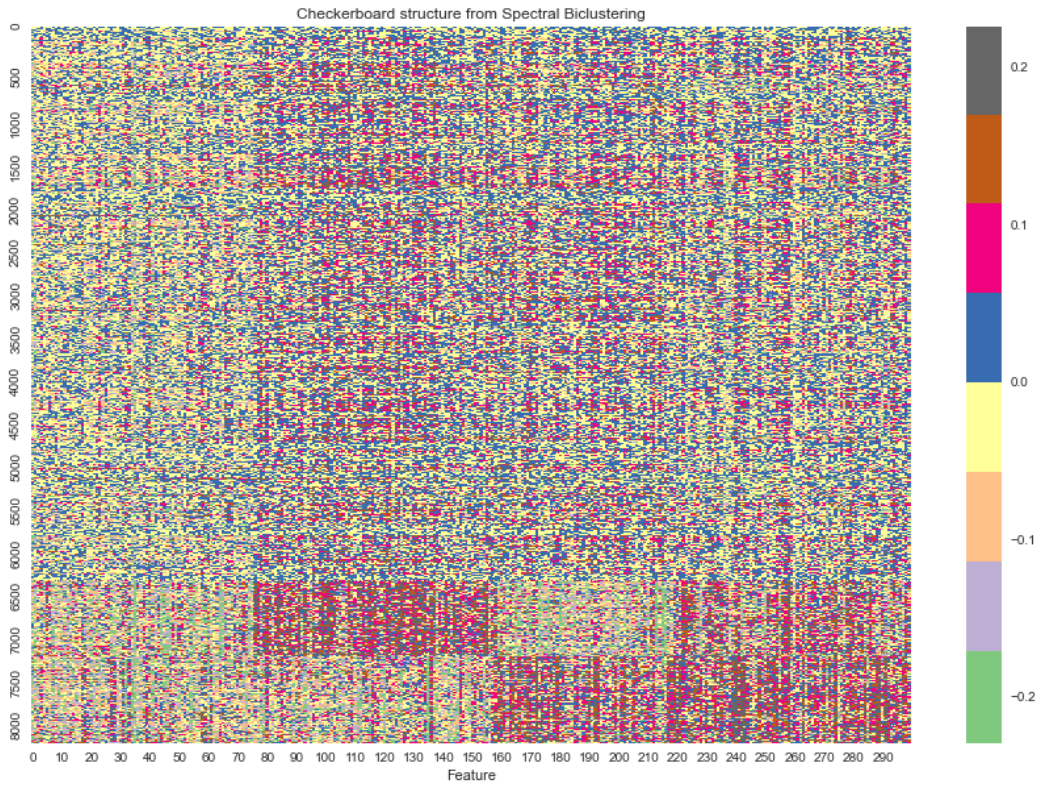
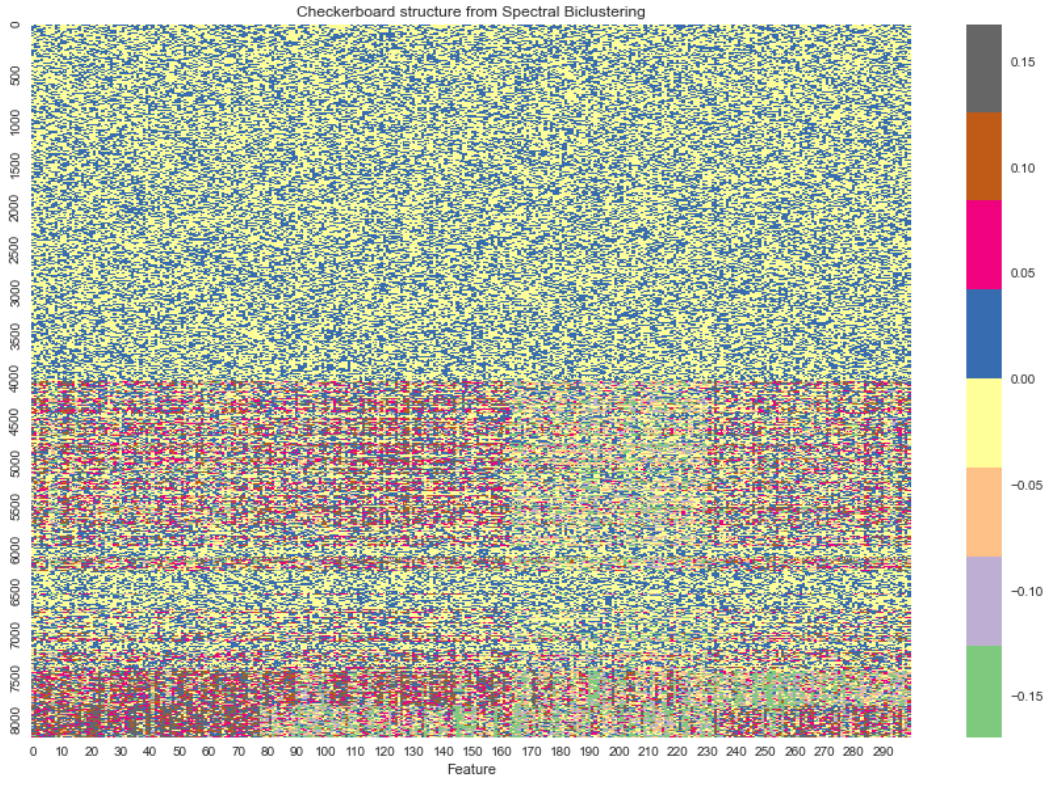
Press questions	
Ms. Psaki's answers	

Figure 8 *Checkerboard visualizations of the press questions and Ms. Psaki's answers.* Both sets of data were vectorized together at 300 dimensions with the configuration of 3 row clusters and 4 column clusters which yielded the highest consensus score.

Figure 8 showed the sorted checkerboard visualizations of both the press questions and Ms. Psaki's answers at such biclustering configurations. The checkerboard patterns were visible. The data in each row cluster were examined, and unfortunately the topics of the clusters were not apparent.

5.4 VADER Sentiment Analysis

We applied VADER sentiment analysis to the original Q&A to obtain a compound sentiment score for each of the questions and each of the answers. The compound sentiment score ranges from [-1,1], with -1 being most negative and 1 being most positive. After scoring each of the questions and answers, we aggregated the sentiment scores for the press and Ms. Psaki to obtain their weekly average sentiment scores.

Figure 9 showed the weekly sentiment scores for the press and Ms. Psaki overlaying the weekly presidential job approval/disapproval ratings. The press sentiment had been largely stable throughout the Administration and remained on the slightly positive side, while Ms. Psaki's sentiment had been always higher than that of the press yet dropping over time.

Figure 10 showed a correlation matrix between the press and Ms. Psaki's sentiments, the weekly presidential job ratings, and the week-on-week job rating changes. The press sentiment showed very mild correlation to Ms. Psaki's sentiment. The press sentiment did not show much correlation with the job approval ratings and their movement, however, Ms. Psaki's sentiment showed a moderately positive correlation to the job approval rating and a moderately negative correlation to the job disapproval rating. It is worth noting that her sentiment did not show correlation with the week-on-week job rating movements. The interpretation of this will be further discussed in Section 6.3.

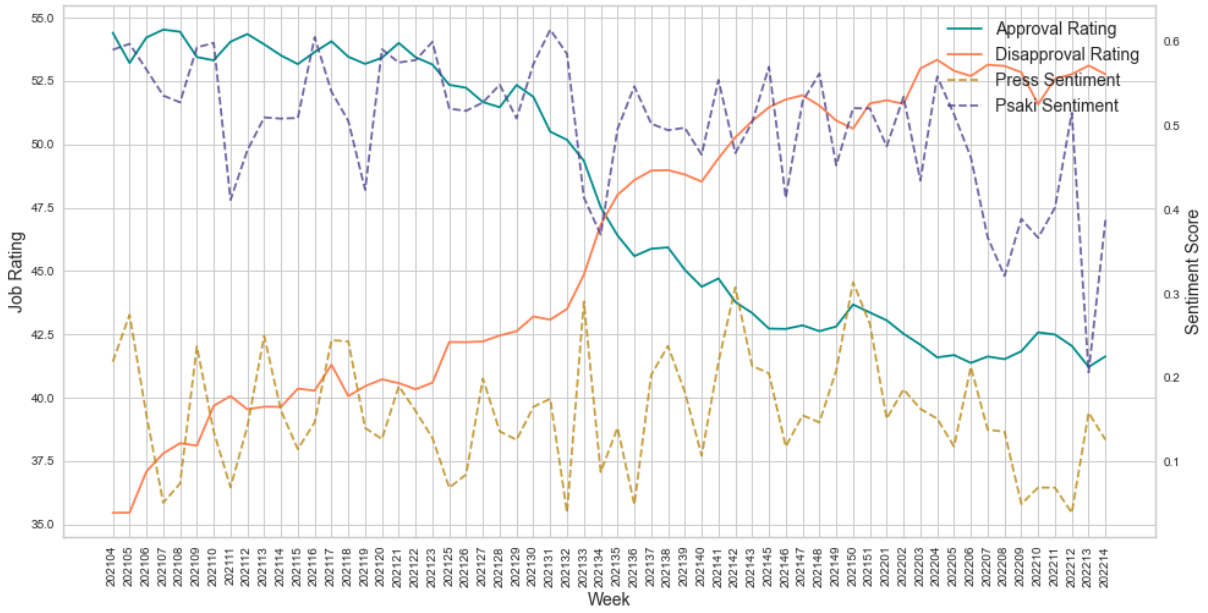


Figure 9 *Weekly sentiment scores for the press and Ms. Psaki overlaying the weekly presidential job approval/disapproval ratings.* While the sentiments from both the press and Ms. Psaki had been positive, the press sentiment had remained relatively stable throughout the Administration while Ms. Psaki's sentiment decreased. The decrease in her sentiment coincided with the drop in presidential job approval ratings.

With the sentiment scores for both the press and Ms. Psaki, we calculated the differences between the two on each Q&A pairs and applied our method from Section 4.1 to extract top proper nouns to gain insights into the topics whereas the press and Ms. Psaki shared closer and further apart sentiments. Figure 11 showed the word clouds of top proper nouns for when the sentiment differences were below the 25th percentile of the distribution, and for when the sentiment differences were above the 75th percentile. It appeared that the press and Ms. Psaki shared closer sentiment on the topic of the war in Ukraine and Russia, and differed on the topic of Afghanistan. "Covid", one of the top proper nouns in our corpus, was notably absent in both polarities. "Peter", upon inspection, refers to Peter Doocy, the current White House Correspondent from Fox News. The word was in the transcript when Ms. Psaki replied to Mr.

Doocy. Given the relationship between the White House and Fox News, it was surprising to see Mr. Doocy's name showed up as a top word when their sentiments were similar.

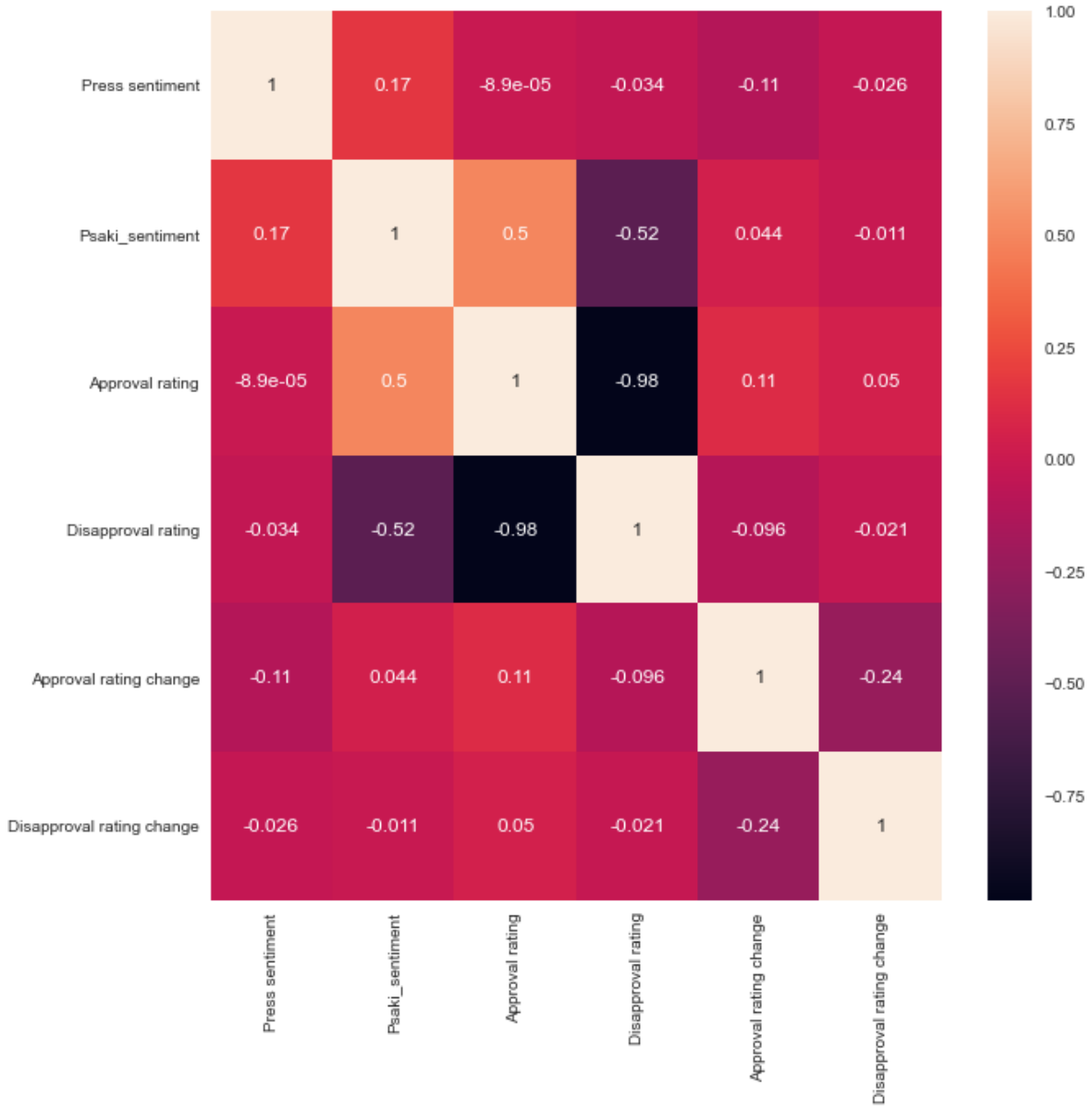


Figure 10 *Correlation matrix of the press and Ms. Psaki's sentiments, the weekly presidential job ratings, and the week-on-week job rating changes.* Ms. Psaki's sentiment correlated moderately with the presidential job approval/disapproval ratings.

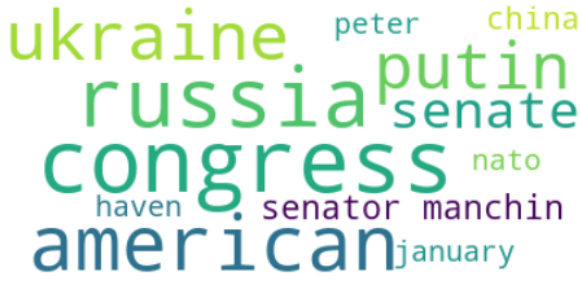
Sentiment score differences < 0.28 (below 25th percentile of the distribution)	Sentiment score differences > 0.41 (above 75th percentile of the distribution)
	

Figure 11 *Top proper nouns in the time when the press and Ms. Psaki shared closer (left) and further apart (right) sentiments.* The words were indicative of the topics of discussion at the time.

6. Analysis and Interpretation

We obtained many insights during our exploration of the dataset using POS and OOV words extraction, unsupervised clustering methods and sentiment analysis. Below we discussed and interpreted the findings.

6.1 The Usage of POS and OOV Words Extraction

The use of proper nouns and OOV words as a means for topic extraction stood out as a key finding of this study. These words are particularly meaningful in the study of press briefing and the world of politics. Many names of countries, people, and organizations such as "Afghanistan", "Fauci", and "NATO" are strongly indicative of the topics by themselves. While POS/OOV extraction is not always useful in general document categorization, it has been immensely useful in our specific case to help human analysts interpret quickly what was being discussed. Using this as a background tool to integrate with other datasets such as the

presidential job approval ratings, we were able to glimpse at the driving force behind rating movements.

6.2 Unsupervised Clustering Methods

We approached the clustering of our dataset using k-means clustering and spectral biclustering. In k-means clustering, we observed that vectorizing to higher dimensional vectors allowed a greater level of details about the data to be retained, and was advantageous to the clustering effort. The trade off in using higher dimensional vectors would be the computing cost, although for this study the cost remained minimal throughout.

As we extracted more key details and discarded noises from low-meaning words by lemmatization and proper nouns tagging, the silhouette scores of our k-means clustering climbed higher indicating clusters with clearer definition. The trade off in this was that some of the Q&A exchanges without lemmas or proper nouns were lost during the discarding, and we risked biasing our analytical judgment towards topics that were richer in proper nouns. K-means clustering operates on the assumption that there is a centroid location for each cluster and the clustering is based on the distance from the centroid. This assumption is not always true especially for non-convex problems where the data clusters may exist in multiple dimensions. Our Q&A dataset falls into this category and other clustering methods could be more suitable.

Spectral biclustering offered an additional dimension to cluster our data with both the data documents (rows) and vectorized features (columns), which was a more suitable clustering method. The clustering of vectorized features allowed us a means to factor in the finer details of each Q&A. We used consensus score to explore the coherence between the press and Ms. Psaki in our paired data. Based on the premise that both sides were likely to be talking about the same

topic in a given exchange, we leverage this as a way to gauge the effectiveness of our unsupervised clustering.

In both k-means clustering and spectral biclustering, the meaning of the resulting clusters were not immediately clear. This was owing to our process of vectorizing the Q&A using doc2vec, which involved vectorizing a document entry as a whole versus other methods such as TF-IDF Vectorizer which were based on individual words with top frequencies. Doc2vec produced vectors for the Q&A which were less interpretable to humans, and that posed difficulties for our interpretation of the clustering outcomes.

6.3 VADER Sentiment Analysis

Our application of the VADER sentiment analysis tool has proven to be an interesting study. VADER was originally built to gear towards gauging sentiment in social media. It is based on a lexicon while leveraging sentiment-laden words such as “great” or “violence”, degree modifying words such as “very”, punctuations, slangs and emojis for assessing sentiment. While some of these clues such as punctuations and emojis were not applicable in a White House press briefing transcript, the spoken language and occasional banters between the press and Ms. Psaki bore similarity to the casual language in social media.

Despite the fact that some individual exchanges between the press and Ms. Psaki can get very positive or very negative, the weekly average sentiment from both sides remained positive throughout the Administration. Civility and positive outlook are certainly traits of a constructive discussion space, and the weekly sentiments obtained in this study pointed to such constructive conversations.

Ms. Psaki's sentiment was shown to be decreasing throughout the study timeframe albeit remaining on the positive side. One of the possible influences could be the nature of the news growing more negative over the course of time. After all, our timeframe of the year of 2021 and early 2022 had been turbulent with heated debates domestically in the US and violent wars on an international level, all the while the country was operating under an overarching theme of the COVID-19 pandemic. The conjecture that Ms. Psaki's dropping sentiment due to externalities in the news was unlikely though as the press sentiment appeared to have remained relatively stable in the same timeframe. A moderate correlation was seen between Ms. Psaki's sentiment and the presidential job approval ratings. Both were shown to be decreasing over our course of study. Correlation does not necessarily imply causation, and we did not find evidence that Ms. Psaki's sentiment drop was caused by the drop in approval ratings. In fact, the week-on-week approval rating change did not correlate with Ms. Psaki's sentiment. Based on the findings in this study, we believe that there are other contributing factors to the downtrend of Ms. Psaki's sentiment.

The uniqueness of our dataset was that the Q&A came in pairs. With VADER we gained the ability to score sentiment by each question and answer, and it offered us an opportunity to assess the interaction between the press and Ms. Psaki. We extracted the topics where both sides shared similar or very different sentiments through proper nouns tagging, and we arrived at insights resembling the ones we learned in Section 5.1. The war between Ukraine and Russia was a topic where both sides shared similar sentiment, and was also the driving topic in the weeks when the presidential job approval ratings were up and disapproval ratings were down. The topic of Afghanistan was the opposite for job ratings, and was also a topic where the press and Ms. Psaki showing differing sentiments. The press were serving as representatives of the public in the White House Briefing Room, while Ms. Psaki was of the Administration. One may

infer that when the press and Ms. Psaki shared similar sentiments, it would be when the public was more approving of the positions and actions of the Administration.

7. Conclusions

An explorative study was conducted to better understand public opinion through analyzing White House press briefing transcripts using NLP techniques, unsupervised clustering methods, and sentiment analysis. We curated a dataset of Q&A exchange pairs at press briefings under the Biden Administration since its inauguration in January 2021 to April 2022 using web scraping and text processing methods. We obtained key topics of public interest over the study timeframe via the extraction of proper nouns using spaCy, an NLP tool for understanding language semantics. The Q&A dataset was vectorized using doc2vec and we applied unsupervised clustering methods including k-means clustering and spectral biclustering to cluster the data. We learned that vectors of higher dimensions retained more detail from the data which was beneficial for clustering. Spectral biclustering offered an extra dimension to the clustering itself, and was explored as a way to gauge coherence within the Q&A paired data. The sentiments from both sides of the Q&A were scored using VADER sentiment analysis. The Q&A sentiment differences were found to point to topics that were driving presidential job approval ratings. This work laid the foundation for extracting public opinion with the objectivity of science via the lens of White House press briefings, which are a main communication channel between the public and the Administration. It is our hope to aid our government to better serve our people by providing an innovative way to understand public opinion.

8. Directions for Future Work

Through the exercise of looking into top words in relation to presidential job approval rating, we learned that the proper nouns and OOV words extraction is a powerful topic identification tool for gaining additional insights on the approval rating dataset quickly. It would be of interest to apply the same technique on other datasets such as the S&P 500 index to observe the influence that news topics have on various data movements.

Unsupervised clustering of our dataset could be useful for sorting the Q&A exchanges into topics and allow us to analyze for public opinion further, however our vectorization approach could be experimented further with other methods such as TF-IDF Vectorizer which could give more human-interpretable meaning to our vectors. Non-convex based clustering methods such as spectral biclustering had proven to be useful, and future work should explore similar methods for better results.

With the sentiment analysis in this study, it would be of interest to continue scoring of the sentiments at White House press briefings. Seeing that the sentiment differences between the press and Ms. Psaki might be indicative of the topics earning public approval, a longer-term exploration in this direction could prove beneficial for understanding public opinion better. It could be a useful feature to have for building machine learning models to predict presidential job approval ratings.

At the time of this writing, Ms. Psaki has left the position and Ms. Karine Jean-Pierre has stepped in as the next White House Press Secretary. With a new Press Secretary, the interactions in the Briefing Room will likely change and could offer new insights on our understanding of US politics and public opinion in this country. A continuation of this work would be of interest to many in the political world.

9. Acknowledgements

I take this opportunity to express gratitude to Dr. Thomas W. Miller and Dr. Alianna J. Maren of Northwestern University for their guidance during the exploration of the topic. I would also like to thank my colleagues Ms. Riddhi Gandhi and Ms. Prasanthi Desiraju for their input to the writing of this work.

Appendix

Appendix 1 An example of press briefing listing on the White House website

Appendix 2 An example of a press briefing transcript on the White House website

THE WHITE HOUSE

Administration Priorities COVID Plan Briefing Room Español MENU

BRIEFING ROOM

Press Briefing by Press Secretary Jen Psaki, April 20, 2022

APRIL 20, 2022 • PRESS BRIEFINGS

Share

James S. Brady Press Briefing Room

3:05 P.M. EDT

MS. PSAKI: Hi, everyone. Hello. Okay, a couple of items for all of you at the top.

As you know, the President is headed to the West Coast tomorrow, and I wanted to give you a quick preview of his first trip to Oregon and Washington. He will highlight the historic economic growth and nearly 8 million jobs created as a result of his and congressional Democrats' actions, including the American Rescue Plan and Bipartisan Infrastructure Law, and his work to lower costs.

He will visit Portland International Airport tomorrow to highlight critical investments to ensure stronger, more resilient infrastructure, such as an earthquake-resilient runway at the Portland Airport, and to help lower costs on everyday items by ensuring goods can move faster and more efficiently.

Appendix 3 An extraction from a press briefing transcript on the White House website

Q Jen, a Ukrainian commander in Mariupol is saying that his forces and civilians in the area may only have hours to make it out alive, and he has been making this direct plea to the President for the U.S. to somehow be involved in extracting people from the area. Do you know if that's a request that the President himself is aware of? And can you just walk us through — can the U.S. be directly involved at this point in making that happen or helping to make that happen?

MS. PSAKI: Sure. We have certainly seen these — these cries for help and these asks for help. What — we certainly urge the Russian government to do the right thing: guarantee safe passage for any civilians or others who wish to leave the city. We also encourage them to allow deliveries of humanitarian aid, such as food and medicine and safe passage for volunteers to help people in need. But I don't have any additional updates from here.

Q And quickly, on the President and whether he might visit Ukraine, you've obviously said a couple of times: Currently, there are no plans to do that. Yesterday, when he was asked "Are you going?," he said he wasn't sure. Is a trip completely off the table for now for him?

MS. PSAKI: Nothing has changed in our assessment. I would also note, as I've said in here before, if anyone were to visit Ukraine from the United States, it's not something we would announce in advance, confirm in advance, give details on who, if, and when because of security reasons.

And that's something I think anyone who's traveled to a war zone with a President or a Vice President or anyone is quite familiar with that policy.

Bibliography

Bordalo, Pedro, David Yang, and Zilibotti Doepke. 2021. "Misinformation on social media." VOX, CEPR Policy Portal. https://voxeu.org/article/misinformation-social-media.

Bush White House Archives. n.d. "Press Briefing Archives." Bush White House Archives. https://georgewbush-whitehouse.archives.gov/news/briefings/.

Byler, David. 2021. "Opinion | Polling is broken. No one knows how to fix it." *The Washington Post*, July 22, 2021. https://www.washingtonpost.com/opinions/2021/07/22/polling-is-broken-no-one-knows-how-fix-it/.

FiveThirtyEight. 2022. "How Popular Is Joe Biden?" FiveThirtyEight. https://projects.fivethirtyeight.com/biden-approval-rating/.

Haberman, Maggie, Jeremy W. Peters, and Elaina Plott. 2020. "260000 Words, Full of Self-Praise, From Trump on the Virus (Published 2020)." *The New York Times*, April 26, 2020. https://www.nytimes.com/interactive/2020/04/26/us/politics/trump-coronavirus-briefings-analyzed.html.

Hutto, C., and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1 (May). https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

jjgoings. 2020. "What does Trump talk about when he talks about COVID-19? (A natural language processing exploration.)." GitHub. https://github.com/jjgoings/trump-covid-briefings.

Kluger, Yuval, Ronen Basri, Joseph T. Chang, and Mark Gerstein. 2003. "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions." *Genome Research* 13, no. 4 (April): 703-716. https://genome.cshlp.org/content/13/4/703.

Le, Quoc V., and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *arXiv:1405.4053 [cs.CL]*, (May). https://arxiv.org/abs/1405.4053.

MacQueen, James. 1967. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1 (14): 281-297.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv:1301.3781 [cs.CL]*, (September). https://arxiv.org/abs/1301.3781.

Mohamed, Sakhri. 2021. "Importance of Public Opinion." Algerian Encyclopedia of Political and Strategic Studies. https://www.politics-dz.com/en/importance-of-public-opinion/.

National Archive. n.d. "Archived Presidential White House Websites | National Archives." National Archives. https://www.archives.gov/presidential-libraries/archived-websites.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems* 14 (January): 849–856. https://dl.acm.org/doi/10.5555/2980539.2980649.

Obama White House Archives. n.d. "Press Briefings | whitehouse.gov." Obama White House Archives. https://obamawhitehouse.archives.gov/briefing-room/press-briefings.

PBS. n.d. "White House Press Briefings." PBS. Accessed April 22, 2022. https://www.pbs.org/newshour/tag/white-house-press-briefings.

Politico. n.d. “White House press briefing- POLITICO.” Politico. Accessed April 22, 2022. https://www.politico.com/news/white-house-press-briefing.

Rakich, Nathaniel. 2021. “How We're Tracking Joe Biden's Approval Rating.” FiveThirtyEight. https://fivethirtyeight.com/features/how-were-tracking-joe-bidens-approval-rating/.

Rasmussen Reports. n.d. “Daily Presidential Tracking Poll.” Rasmussen Reports. Accessed April 24, 2022. https://www.rasmussenreports.com/public_content/politics/biden_administration/prez_track_apr22.

Reuters. 2022. “Biden approval polling tracker.” Reuters. https://graphics.reuters.com/USA-BIDEN/POLL/nmopagnqapa/.

Scikit-learn. n.d. “sklearn.cluster.SpectralClustering — scikit-learn 1.0.2 documentation.” Scikit-learn. Accessed April 29, 2022. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html.

Shaffer, Kris. 2017. “Data mining whitehouse.gov. What has the Trump administration changed on whitehouse.gov?” Medium. https://medium.com/@krisshaffer/data-mining-whitehouse-gov-5ee12b9709a5.

spaCy. n.d. “spaCy · Industrial-strength Natural Language Processing in Python.” spaCy. Accessed April 22, 2022. https://spacy.io.

Stimson, James A. 1999. *Public Opinion in America: Moods, Cycles, and Swings (Transforming American Politics)*. N.p.: Avalon Publishing.

UC Santa Barbara. n.d. “Presidential News Conferences.” The American Presidency Project. Accessed April 21, 2022. https://www.presidency.ucsb.edu/statistics/data/presidential-news-conferences.

The White House. n.d. “Press Briefings Archives.” Accessed April 24, 2022. https://www.whitehouse.gov/briefing-room/press-briefings/.

YouGov. n.d. “President Biden job approval rating.” YouGov. Accessed April 24, 2022. https://today.yougov.com/topics/politics/trackers/president-biden-job-approval-rating.