Shivani Chowdury, Sharon Jepkosgei, Ryan Rynbrandt, Liberty Smith
EPPS 6302 - Methods of Data Collection and Production
Dr. Karl Ho
November 26, 2023

News Coverage of Asian-American Pacific Islanders

Introduction:

Asian American Pacific Islander (AAPI) communities have long been underrepresented and misrepresented in the media, including in national news outlets, where AAPI are often portrayed in stereotypical ways or as a monolith. This paper examines how AAPI communities are portrayed in two major national newspapers, the New York Times (NYT) and the Wall Street Journal (WSJ). The study focuses on two main research questions: (1) Does the partisanship of national news outlets influence how they portray AAPI? and (2) Did the portrayal of AAPI in national news outlets change after the COVID-19 pandemic?

To address these questions, we collected a corpus of 381 articles about AAPI from the NYT and WSJ published in 2019 and 2021. We employed various methods to analyze the data, including topic modeling, bag of words and word frequency analysis, word ratio analysis, and sentiment analysis. Our findings provide evidence that news outlet partisanship does influence how AAPI are portrayed and that the portrayal of AAPI changed after the COVID-19 pandemic.

Review of the Literature

AAPI are the fastest-growing ethnic group in the United States, accounting for 7% of the population and growing by 81% between 2000 and 2020 (U.S. Census Bureau, 2022). As with other minorities, AAPI have faced discrimination and stereotyping throughout American history. Most recently, the COVID pandemic led to an increase in anti-Asian hate crimes in the US, raising concerns about how AAPI are portrayed in the media. Since media outlets play a powerful role in shaping views of the world, it is important to consider how different media outlets portray AAPI, and how this may influence what people think and believe about them.

Previous research has shown that the partisanship of news outlets can influence their coverage of a variety of topics, including race and ethnicity. For example, one study found that conservative news outlets were more likely to portray African Americans as criminals and liberal news outlets were more likely to portray them as victims of discrimination (Iyengar & Kinder, 1987). Tsai et al (2020) found consuming news from left-leaning and neutral outlets during the pandemic to be associated with less prejudicial attitudes toward Asians.

AAPI are often stereotyped in the media as being hardworking, intelligent, and successful. Though such "model minority" stereotypes highlight positive attributes and achievements, they can be harmful in a variety of ways. Being stereotyped as high-achievers results in perceptions of AAPI as having fewer needs than other groups or being overlooked for opportunities and support. Additionally, it can lead to AAPI feeling pressure to live up to unrealistic expectations and to avoid seeking help when they need it. (Sue et al, 2007). After the pandemic, more clearly negative stereotypes began to arise. One study analyzed a sample of 1,157 news articles from five major US news outlets and found that AAPI were more likely to be portrayed as dirty, diseased, and dangerous to public health in the media during the pandemic and were more likely to be blamed for the pandemic (Jang et al, 2021). Such coverage has been shown to affect people's attitudes. Santia et al (2022) found that exposure to an anti-Asian news story during the pandemic negatively affected attitudes toward the depicted group.

Our study adds to the existing literature in two important ways. First, whereas previous research looked at the media as a monolithic entity, we seek to explain the influence of media outlets' partisan leanings in their portrayals of AAPI. Second, we seek to explain the impact of the COVID pandemic on changes in media portrayals of AAPI.

Research Questions:

Main research questions:

1. Does the partisanship of national news outlets influence how they portray AAPI?
2. Did portrayals of AAPI in national news outlets change after the COVID-19 pandemic?

Additional/secondary research question:

3. To what extent do either liberal or conservative outlets perpetuate stereotypes surrounding AAPI?

Hypotheses:

First hypotheses:

H01: News outlet partisanship has no influence on their portrayal of AAPI
HA1: Conservative and liberal news outlets will portray AAPI differently

Second hypotheses:

H02: Portrayals of AAPI did not change from pre-pandemic to post-pandemic
HA2: Portrayals of AAPI changed from pre-pandemic to post-pandemic

Additional/Secondary hypotheses:

H03: There will be no difference between liberal and conservative outlets in perpetuation of stereotypes about AAPI
HA3: Conservative news outlets will be more likely to perpetuate stereotypes about AAPI than liberal news outlets.

Data

To test our hypotheses, we collected a corpus of 381 articles about AAPIs from two major U.S. newspapers, the Wall Street Journal (WSJ) and New York Times (NYT) published in 2019 and in 2021 using the ProQuest LexisNexis Academic database.

The choice of outlets allows us to investigate differences between conservative and liberal media outlets, as WSJ is widely regarded as a leading conservative news outlet and NYT is widely regarded as a leading liberal news outlet. Since WSJ is a center-right publication and NYT is a center-left publication and both are recognized as having a greater focus on accuracy in reporting than on analysis or ideological agenda-pushing, the choice of these outlets provides a strong test of the hypothesis that there will be an ideological difference (Ad Fontes, 2023). Both are globally influential major newspapers of record with massive readership and long-established reputations for excellence in reporting and first-hand coverage. Thus, we would expect less divergence in portrayals of AAPI in these sources than we would in a comparison of more extreme and opinionated sources or in a comparison of very different types of outlets.

We also chose NYT and WSJ due to their extensive influence in the American media context. Not only do they have large readerships, they are often cited as sources of news and information by other media outlets. When they break a story, it is often picked up by other news outlets. This can help to set the agenda for what other media outlets cover and can also influence how they frame stories (Baum & Peters, 2013). Their editorials also have been found to shape public discourse and influence the way other media outlets cover certain issues (Farrell, 2007). In addition, both outlets have numerous syndication deals that result in their articles and op-eds being published in other newspapers and websites (Napoli, 2011). Thus what is reported in these two outlets is representative of more than just what is contained within their respective pages.

To investigate the difference between the pre-pandemic and post-pandemic media environment, we chose one collection of articles from 2019 and another from 2021. Though the novel coronavirus emerged in China in October of 2019, it was not yet a major agenda item for the American media and public, and as such 2019 is a valid representation of a “pre-pandemic” public reality. The choice of years deliberately excludes 2020, which included several months at the beginning of the year that may be considered “pre-pandemic” in the sense described above and several subsequent months where much of the future of the pandemic and the governmental responses to it were unclear. In contrast, by 2021 the pandemic was a fully established reality in American life. As a result, 2021 is more fully representative of a “post-pandemic” public reality.

Methodology

Here we elucidate the processes and rationales behind the topic modeling, bag of words, word ratios, sentiment analysis, and bi-gram analyses in this study, highlighting their significance in uncovering thematic trends in media portrayal. The graphics generated from this analysis will visually articulate these findings, demonstrating how media narratives around AAPI communities have evolved. Each methodological step contributed to a detailed and nuanced analysis of language used in media portrayals of AAPI. By systematically categorizing, filtering, and analyzing the text data, we uncover implicit biases, thematic focuses, and potential shifts in portrayal across different time periods and media outlets.

Preprocessing

Preprocessing the text data involved several steps. Initially, articles were converted into plain text format for ease of processing. Given the focus on AAPI-related content, we identified a specific set of keywords and phrases related to AAPI, such as 'Asian American,' 'Pacific Islander,' and 'AAPI.' This word set was instrumental in filtering and categorizing the text data.

Bag of Words and Word Frequency Analysis

A primary method of analysis was the bag of words approach, a common technique in NLP used to simplify text data by converting it into a set of words. In this context, the bag of words model served two main purposes: to identify the frequency of specific words related to AAPI within the articles and to facilitate comparative analysis across time frames and media outlets. The Python script for processing the text data was instrumental in this regard. It included functions for categorizing sentences based on the presence of AAPI-related words, counting word frequencies, and identifying proper nouns that might skew analysis. The script further refined the data by removing stop words and filtering out proper nouns not central to analysis.

The process began with the establishment of a specific wordset, AAPI_wordset, which includes terms like 'Asian American,' 'Asian-american,' 'Asian,' 'AAPI,' 'Pacific Islander,' and 'Pacific-islander.' This wordset was crucial for identifying and focusing on content relevant to the AAPI community within the corpus. The text of each article was first split into sentences using a tokenizer. This division into sentences was essential for subsequent categorization and analysis at the sentence level, which provided a more nuanced understanding of the context in which AAPI-related terms were used. Each sentence was processed to remove standard English stop words and punctuation. This step reduced the data to its meaningful components, eliminating common words that do not contribute significantly to the analysis of AAPI portrayal. Each sentence was categorized based on the presence of words from the AAPI_wordset. This allowed for targeted analysis of sentences that specifically mention or relate to the AAPI community.

For each sentence categorized under the AAPI category, the frequency of each word was counted. This counting was pivotal in understanding the prevalence and emphasis of certain words in the context of AAPI discussions. The script also included a function to identify proper nouns and to determine whether they should be counted based on their case (uppercase or lowercase). Proper nouns with a high proportion of uppercase occurrences were filtered out. This step focused analysis on common nouns and adjectives, providing a clearer picture of the thematic content rather than individual names or specific entities. The most frequently occurring words in the AAPI category were compiled into a list, from which the top 1000 words were selected. This list was further refined by removing proper nouns and AAPI-specific terms, resulting in a set of 'common words' that were used disproportionately in AAPI-related contexts.

A notable aspect of the methodology was the focus on 'common words' within AAPI-related sentences. These common words were identified by extracting the top 1000 words from the AAPI category and then filtering out proper nouns and AAPI-specific terms. This approach allowed for identification of words that were disproportionately used in the context of AAPI discussions, providing insights into the thematic focus and possible biases in the articles.

Through the application of NLP techniques like bag of words and word frequency analysis, this study offers a methodical approach to understanding media portrayals of AAPI communities. The subsequent analysis, rooted in these methodologies, provides empirical insights into the influence of partisanship in media narratives and the evolution of these narratives in the context of significant global events like the COVID-19 pandemic.

The culmination of the word frequency analysis is presented in a comprehensive word frequency table. This table displays the top 50 words associated with AAPI topics within the dataset. The frequency of each word is indicative of its prominence in the discourse surrounding the AAPI community in the selected articles. The tables serve as a cornerstone for our analysis, enabling a nuanced understanding of the thematic elements prevalent in the media portrayal of AAPI communities. It also provides a basis for further statistical analysis and interpretation in the context of our research questions. This approach is particularly effective in highlighting subtle differences in language use that may reflect broader societal and editorial perspectives. The resulting word frequency table, derived from this meticulous process, serves as a tangible output of the analysis, encapsulating key aspects of the media's portrayal of AAPI communities.

Word Ratio Analysis

Building on the bag of words approach described previously, this section of the paper explores a novel method of analyzing media representation: word ratio analysis. This method involves comparing the frequency of specific words in articles from the New York Times (NYT) and the Wall Street Journal (WSJ), two outlets with differing editorial leanings. By calculating word ratios, this analysis aims to uncover potential biases and thematic differences in how these outlets portrayed AAPI communities, particularly across two key periods, 2019 and 2021.

The core of this analysis lies in the calculation of word ratios. This involved comparing the frequency of each word in AAPI-related sentences across the two news outlets. The ratio was calculated as follows: for a given word, its frequency in the NYT articles was divided by its frequency in the WSJ articles (and vice versa), adjusted by adding 1 to the denominator to avoid division by zero. This method was applied separately for each outlet and each time period (2019 and 2021), enabling a comparative analysis across both dimensions. The comparative analysis focused on identifying words that were more prevalent in one outlet compared to the other, both in 2019 and 2021. This comparison provides insights into how each outlet might emphasize different aspects of the AAPI experience or narrative.

The results of the word ratio analysis are presented in a series of tables, showing the top words with the highest ratios in each category: words more prevalent in NYT and words more prevalent in WSJ, for both 2019 and 2021. These tables provide a clear visual representation of the differences in language use between the two outlets. While informative, this analysis has limitations. Word ratios provide a quantitative measure but lack qualitative depth; they indicate frequency but not context. Future work could involve more nuanced textual analysis or integrating these findings with qualitative research methods.

Sentiment Analysis

Sentiment analysis, also called opinion mining, is an NLP technique used to determine the neutrality, positivity, and negativity of data. Here we describe how sentiment analysis is used to evaluate the representation of AAPI by NYT and WSJ. One way to analyze the sentiment of a text is to consider the text as a combination of its individual words with the overall sentiment content being the aggregate of the individual word sentiments. Diverse methods and dictionaries exist for appraising opinion or emotion in text. The tidytext package provides access to several sentiment lexicons including the *bing* lexicon which we selected to understand the emotional intent of words from the WSJ and NYT spanning the periods of 2019 and 2021.

Bing is a lexicon structured around single words or unigrams. It contains many English words which are assigned negative or positive sentiment. For sentiment analysis, the data was extracted from the WSJ and NYT articles, followed by conversion of text into tidy text format (one-token-per-row) using unnest_tokens(). The bing lexicon plays a pivotal role in categorizing words from the articles into binary clusters, signifying negative or positive sentiments. The accompanying R scripts outline the stepwise procedure of sentiment analysis. However, the key functions include get_sentiments() which provides access to the specific sentiment lexicons along with accompanying measures for each one, inner_join() used to calculate sentiment in diverse ways, and mutate() which determines net sentiment in each text section.

We analyzed and presented the results of the sentiment analysis using sentiment score plots separated by year and media outlet. A benefit of a data frame with both sentiment and word is the ability to analyze word counts that contribute to each sentiment category. By utilizing count() with arguments of both word and sentiment, we gathered insights on how much each word from each media outlet contributed to each sentiment and how they varied across years.

It is crucial to acknowledge that not every English word is in the lexicons because many English words are for most part neutral. Again, this method does not consider qualifiers before a word, for example “no good” or “not true”; *bing,* as a lexicon-based method, relies solely on unigrams. Nevertheless, for numerous types of writing (like the media articles), sustained sections of sarcasm or negated text are rare, thus minimizing their effect in sentiment analysis.

Topic Modeling Analysis

In the exploration of how NYT and WSJ portray AAPI communities, topic modeling emerges as a pivotal analytical tool. This section of the paper details the application of Latent Dirichlet Allocation (LDA) for topic modeling, focusing on articles from 2019 and 2021. The objective is to uncover thematic structures within the textual data, providing insights into potential shifts in media narratives over time and across different editorial perspectives.

LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups. In text analysis, these unobserved groups are 'topics.' LDA assumes that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA posits that there is a hidden, or 'latent,' topic structure in a corpus of documents. Each document is modeled as a random mixture over latent topics, where each topic is characterized by a distribution over words. Computationally, LDA is a probabilistic model where each document is represented as a distribution over topics, and each topic is represented as a distribution over words. This provides the model with flexibility in terms of topic variety and prevalence. The model's parameters are estimated via a process that iterates between assigning topics to words (based on the current estimate of what topics are about) and updating the estimate of what topics are about (based on the words assigned to them). The 'Dirichlet' in LDA refers to the Dirichlet distributions used in the model. Separate models were created for each dataset: NYT and WSJ, for both 2019 and 2021. The process involved the following steps:

1. **Creation of Document-Term Matrices (DTM)**: DTMs were generated for each set of articles, converting the text data into a matrix format suitable for LDA analysis.
2. **LDA Model Application**: LDA models were applied to each DTM. The initial choice of the number of topics (k = 2) was based on preliminary analysis.
3. **Perplexity Calculation and Topic Number Optimization**: To determine the most suitable number of topics for each model, the perplexity function was applied. Perplexity measures how well a probability distribution predicts a sample and is used here to assess the fit of different LDA models with varying numbers of topics. The optimal number of topics was chosen using the 'elbow method,' a technique often employed in statistical analyses to determine the point beyond which additional variables (in this case, topics) do not significantly improve the model.

Bi-Gram Analysis

We also conducted a bi-gram analysis. This involved processing the text to identify and count two-word phrases (bi-grams). Bi-grams provide contextually richer insights compared to single words, revealing more complex patterns in the data. Bi-gram analysis in this study provides a nuanced layer of understanding to the media portrayal of AAPI communities. By focusing on word pairings, this method uncovers the intricate narrative structures and thematic emphases in NYT and WSJ articles. The insights gained from this analysis are instrumental in addressing our research questions regarding the influence of partisanship and the impact of the COVID-19 pandemic on media narratives. The analysis and visualization of bi-grams serve several purposes in the context of our research. By examining the most frequent bi-grams, we identify prevalent themes or topics in the media portrayal of AAPI. This might include specific issues, events, or descriptors frequently associated with AAPI communities. Bi-grams provide insights into the context in which certain words are used. This is particularly important in media analysis, as it helps reveal the narrative frames or angles from which stories are approached.

The R script outlines a systematic approach to bi-gram analysis. The unnest_tokens function from the tidytext package is used to convert the text data into bi-grams. This process involves splitting the text into pairs of consecutive words, thereby creating bi-grams. The separate function divides each bi-gram into its constituent words (word1 and word2). Following this, the filter function removes common stopwords (using the stopwords function from the tm package) and other irrelevant words. These include specific terms related to the newspaper sources and general terms that are not informative for the analysis. The count function tallies the frequency of each bi-gram, and the results are sorted in descending order of frequency.

Challenges and Solutions in Data Preparation

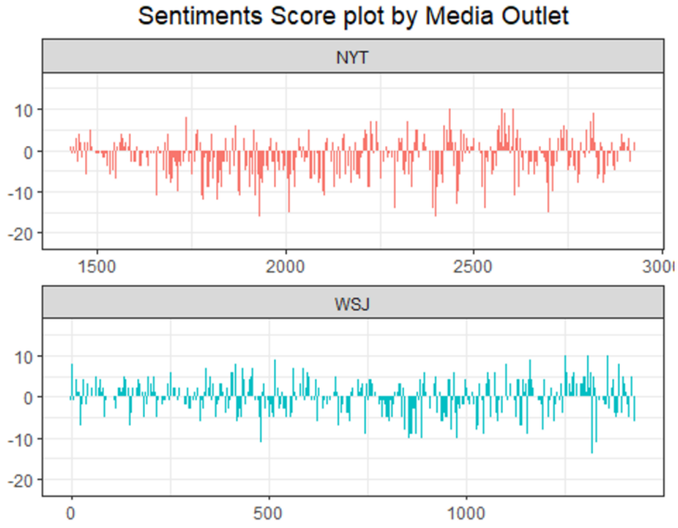
A significant challenge in this process was filtering out the noise created by database-specific terms and formatting used by LexisNexis and ProQuest. By identifying and excluding these terms, the analysis was refined to focus on the meaningful content of the articles. The process of refining the text data for analysis underscored the importance of meticulous data preparation in NLP research. The unique challenge posed by the formatting of the source databases was successfully addressed only in the bi-grams analysis due to time constraints. This was accomplished through the careful selection and inclusion of additional stop words based on its prevalence, or lack thereof, in the dictionaries of the articles.

Results:

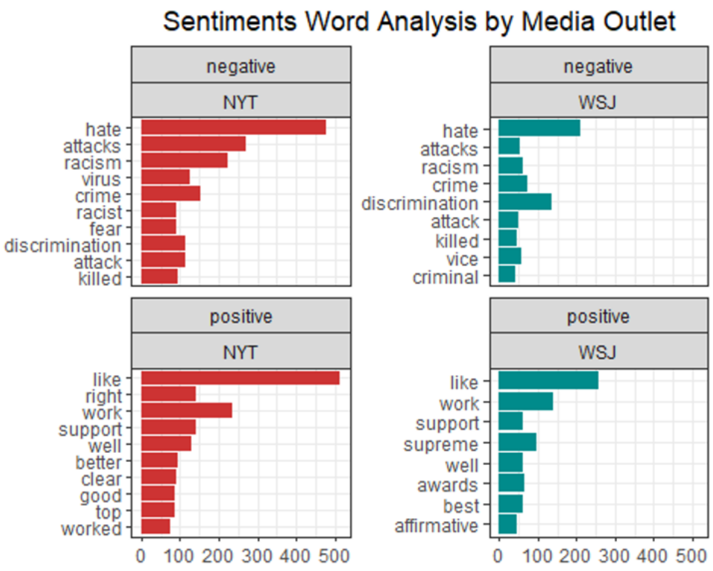
Bing Sentiment Analysis Results:

The initial *bing* sentiment analysis provides some support for our first hypothesis, that there would be a difference between the coverage of AAPI in conservative and liberal national news outlets. Our analysis reveals some differences in the sentiment of coverage of AAPIs in

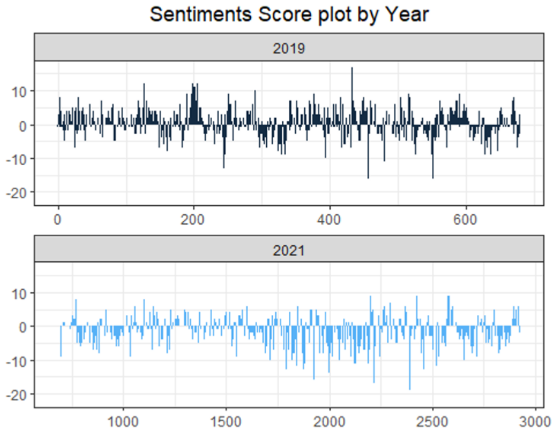
WSJ and NYT from 2019 to 2021. Overall, NYT had a higher proportion and prevalence of negative words than WSJ. It is important to note, however, that the *bing* approach broadly categorizes words as positive or negative. As a result, an increase in negativity does not *necessarily* indicate media negativity towards AAPI as individuals or as a group. Rather, it may indicate coverage of the increased discrimination and violence toward AAPI, as words associated with discrimination and violence would be categorized as negative. The latter interpretation is supported by the dramatic increase in occurrence of words like “hate,” “attacks” and “racism”.



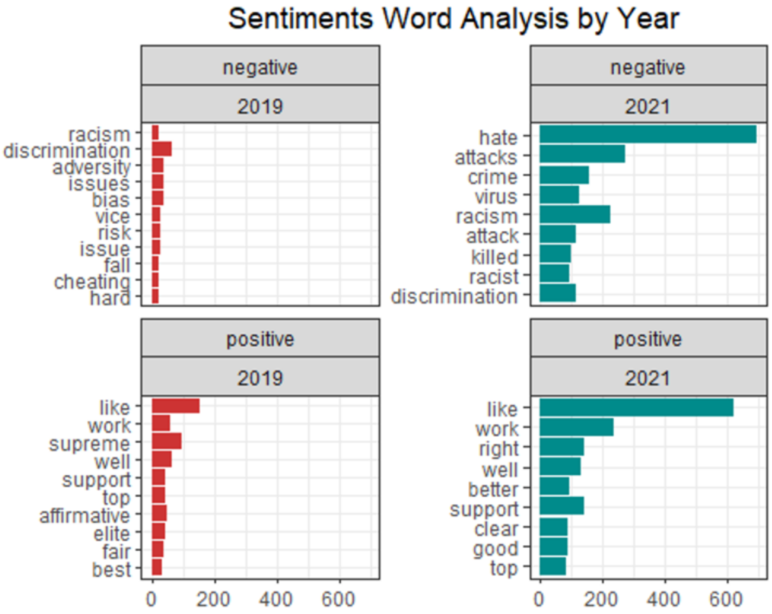
Both the positive and negative words highest in prevalence were remarkably similar between the two outlets, suggesting that the difference between the two outlets was more a difference of prevalence than of content or substance. Thus, we are careful not to overstate the extent of support provided by this analysis for our first hypothesis.



This analysis provides more significant support for our second hypothesis, that national media outlet portrayals of AAPI would change post-pandemic. Our analysis indicates significant differences in the tone of coverage of AAPI pre- and post-pandemic. The proportion and prevalence of negative words was significantly higher in 2021 than it was in 2019.



Again, the selection of words atop the list indicates more coverage of increased violence and discrimination toward AAPI in the post-pandemic US. This interpretation is again supported by the prevalence of words like "hate," "attacks" and "racism". Unlike the test of our first hypothesis, the test of our second hypothesis shows notable change in substance of coverage.

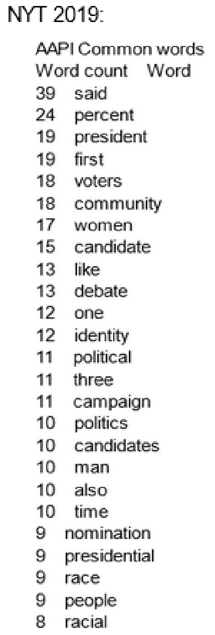


The *bing* sentiment analysis does not provide support for our third hypothesis, that conservative outlets would be more likely to perpetuate stereotypes about AAPI. Here, the sentiment analysis does not reveal words obviously related to stereotyping in either outlet.

Bag of Words and Word Frequency Analysis Results:

Our bag of words / word frequency analysis provides additional support for our first hypothesis, regarding differences between conservative and liberal outlets. But again, the differences observed were more pronounced pre-pandemic than post-pandemic. As in the *bing* analysis, in 2019, WSJ had heavy emphasis on education-related terms with the words “applicants”, “students”, “admissions” and “schools” in the top six most common words. In contrast, NYT’s list of most common words included no words obviously related to education. Instead, the list emphasized politics-related terms, with “president” and“voters” in the top six and other politics-related terms (“candidate”, “debate”, “political”, etc.) throughout the list.

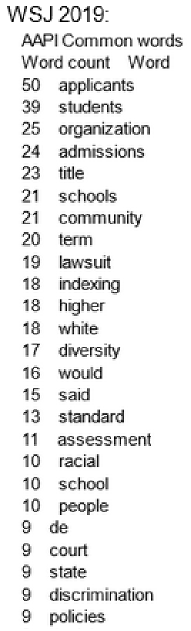
Table 1) Python: Top 25 most used words in 2019 identified after removing stop words, punctuation, filtering specifically for sentences with any reference to AAPI.



NYT 2019:

AAPI Common words

Word count	Word
39	said
24	percent
19	president
19	first
18	voters
18	community
17	women
15	candidate
13	like
13	debate
12	one
12	identity
11	political
11	three
11	campaign
10	politics
10	candidates
10	man
10	also
10	time
9	nomination
9	presidential
9	race
9	people
8	racial



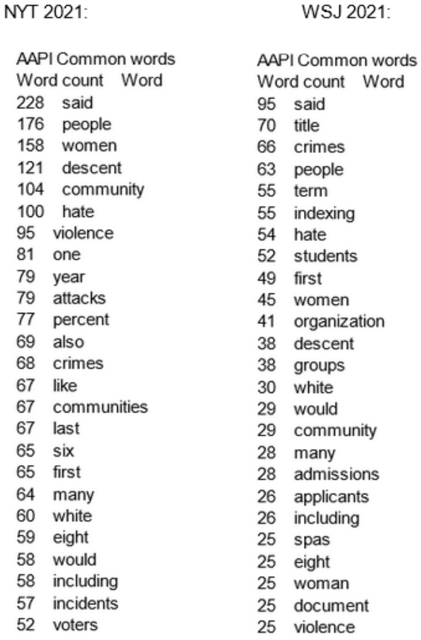
WSJ 2019:

AAPI Common words

Word count	Word
50	applicants
39	students
25	organization
24	admissions
23	title
21	schools
21	community
20	term
19	lawsuit
18	indexing
18	higher
18	white
17	diversity
16	would
15	said
13	standard
11	assessment
10	racial
10	school
10	people
9	de
9	court
9	state
9	discrimination
9	policies

The differences in education emphasis decreased post-pandemic. The 2021 WSJ list still contained more education words than NYT (which had none), but they were a smaller proportion and less frequent than in 2019. Yet differences in emphasis on discrimination, hate and violence were more pronounced between the two outlets in 2021 than in the *bing* analysis, with NYT having a greater prevalence of words like “hate”, “violence” and “attacks”. Thus, our first hypothesis finds additional support in the bag of words / word frequency analysis. Our second hypothesis also received support from this analysis, as the mix of words changed noticeably for both outlets between 2019 and 2021.

Table 2) Python: Top 25 most used words in 2021 identified after removing stop words, punctuation, filtering specifically for sentences with any reference to AAPI.



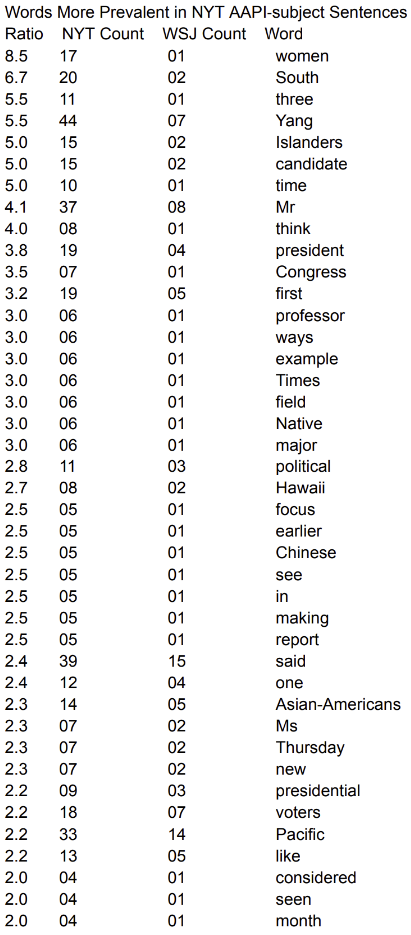
NYT 2021:		WSJ 2021:	
AAPI Common words		AAPI Common words	
Word count	Word	Word count	Word
228	said	95	said
176	people	70	title
158	women	66	crimes
121	descent	63	people
104	community	55	term
100	hate	55	indexing
95	violence	54	hate
81	one	52	students
79	year	49	first
79	attacks	45	women
77	percent	41	organization
69	also	38	descent
68	crimes	38	groups
67	like	30	white
67	communities	29	would
67	last	29	community
65	six	28	many
65	first	28	admissions
64	many	26	applicants
60	white	26	including
59	eight	25	spas
58	would	25	eight
58	including	25	woman
57	incidents	25	document
52	voters	25	violence

The bag of words / word frequency analysis also provides some support for our third hypothesis regarding the perpetuation of stereotypes. The prevalence of education in WSJ strongly suggests the “model minority” stereotype. However, only a qualitative deep dive into the articles can fully confirm that the prevalence of the words corresponds with stereotyping.

Word Ratio Analysis Results:

Word ratio analysis determines the relative prevalence of certain words in NYT articles vs WSJ articles in a specific year relative to the percentage of AAPI-subject sentences to non-subject sentences. This analysis lends support to our first hypothesis regarding differences between conservative and liberal national news outlets. For example, in Table 3 below, the word "women," has a ratio of 8.5, suggesting that it is 8.5 times more prevalent in NYT AAPI-subject sentences compared to those in WSJ. This shows that NYT covers more gender-centric issues related to AAPI in its news coverage than WSJ. Secondly, NYT also appears to have more references to specific geographic references ("South”, "Hawaii”, "Pacific"). Lastly, NYT has more mentions of political terms like "president," "Congress," and "political.", as well as references to specific candidates ("Yang") in the pre-pandemic period. Conversely, words like "Hawaii," and "Chinese" are more evenly distributed between NYT and WSJ.

Table 3. Ratio of words shared between NYT and WSJ in 2019

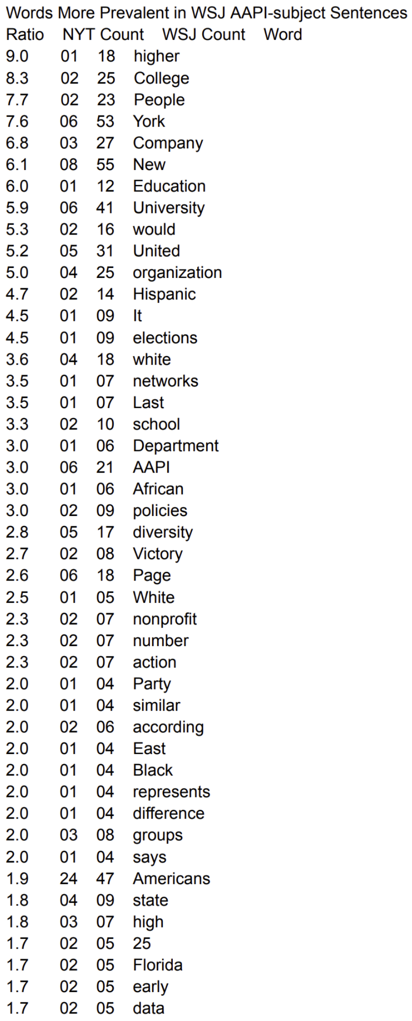


Words More Prevalent in NYT AAPI-subject Sentences

Ratio	NYT Count	WSJ Count	Word
8.5	17	01	women
6.7	20	02	South
5.5	11	01	three
5.5	44	07	Yang
5.0	15	02	Islanders
5.0	15	02	candidate
5.0	10	01	time
4.1	37	08	Mr
4.0	08	01	think
3.8	19	04	president
3.5	07	01	Congress
3.2	19	05	first
3.0	06	01	professor
3.0	06	01	ways
3.0	06	01	example
3.0	06	01	Times
3.0	06	01	field
3.0	06	01	Native
3.0	06	01	major
2.8	11	03	political
2.7	08	02	Hawaii
2.5	05	01	focus
2.5	05	01	earlier
2.5	05	01	Chinese
2.5	05	01	see
2.5	05	01	in
2.5	05	01	making
2.5	05	01	report
2.4	39	15	said
2.4	12	04	one
2.3	14	05	Asian-Americans
2.3	07	02	Ms
2.3	07	02	Thursday
2.3	07	02	new
2.2	09	03	presidential
2.2	18	07	voters
2.2	33	14	Pacific
2.2	13	05	like
2.0	04	01	considered
2.0	04	01	seen
2.0	04	01	month

Next, when it comes to the ratio of words shared between WSJ and NYT in the pre-pandemic period in in 2019, words like "Education," "College," "University”, "Company", “networks”, “organization”, “school”, “diversity” are significantly more prevalent in WSJ compared to NYT, based on the given ratios in Table 4 below.

Table 4. Ratio of words shared between NYT and WSJ in 2019

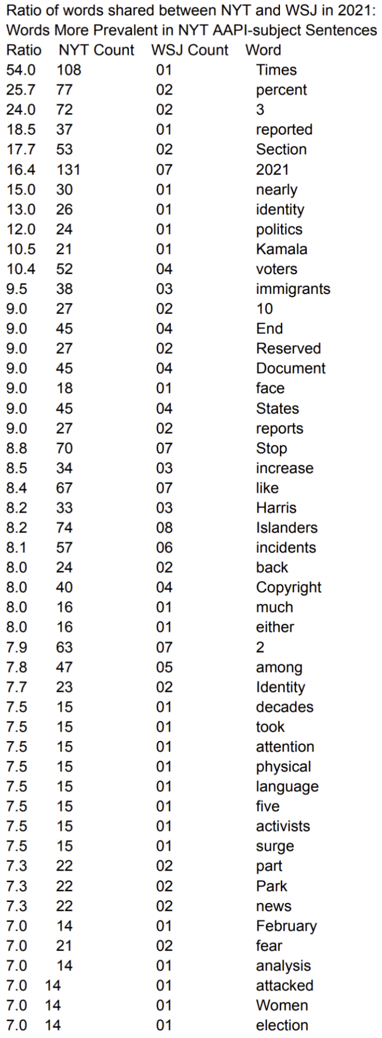


Words More Prevalent in WSJ AAPI-subject Sentences

Ratio	NYT Count	WSJ Count	Word
9.0	01	18	higher
8.3	02	25	College
7.7	02	23	People
7.6	06	53	York
6.8	03	27	Company
6.1	08	55	New
6.0	01	12	Education
5.9	06	41	University
5.3	02	16	would
5.2	05	31	United
5.0	04	25	organization
4.7	02	14	Hispanic
4.5	01	09	It
4.5	01	09	elections
3.6	04	18	white
3.5	01	07	networks
3.5	01	07	Last
3.3	02	10	school
3.0	01	06	Department
3.0	06	21	AAPI
3.0	01	06	African
3.0	02	09	policies
2.8	05	17	diversity
2.7	02	08	Victory
2.6	06	18	Page
2.5	01	05	White
2.3	02	07	nonprofit
2.3	02	07	number
2.3	02	07	action
2.0	01	04	Party
2.0	01	04	similar
2.0	02	06	according
2.0	01	04	East
2.0	01	04	Black
2.0	01	04	represents
2.0	01	04	difference
2.0	03	08	groups
2.0	01	04	says
1.9	24	47	Americans
1.8	04	09	state
1.8	03	07	high
1.7	02	05	25
1.7	02	05	Florida
1.7	02	05	early
1.7	02	05	data

In post-pandemic 2021, COVID-specific terms like "percent," "reported," “February”, "2021" highlight the relevance of the context to the ongoing pandemic in NYT coverage of AAPIs (Table 5 below). Further, among the top words with very high ratios were words like “Identity”, “politics”, “Kamala”, “Harris”, “voters”, “immigrants”, suggesting a distinctive emphasis in NYT coverage, indicating a bigger focus on politics compared to WSJ. Immigration-related terms like "immigrants" and "voters” suggest a connection between AAPI subjects and discussion of immigration policies. The presence of terms like "incidents," "activists," and "attacked" suggests a focus on rising incidents or attacks against the AAPI community in the wake of Covid and activism in support of the AAPI community.

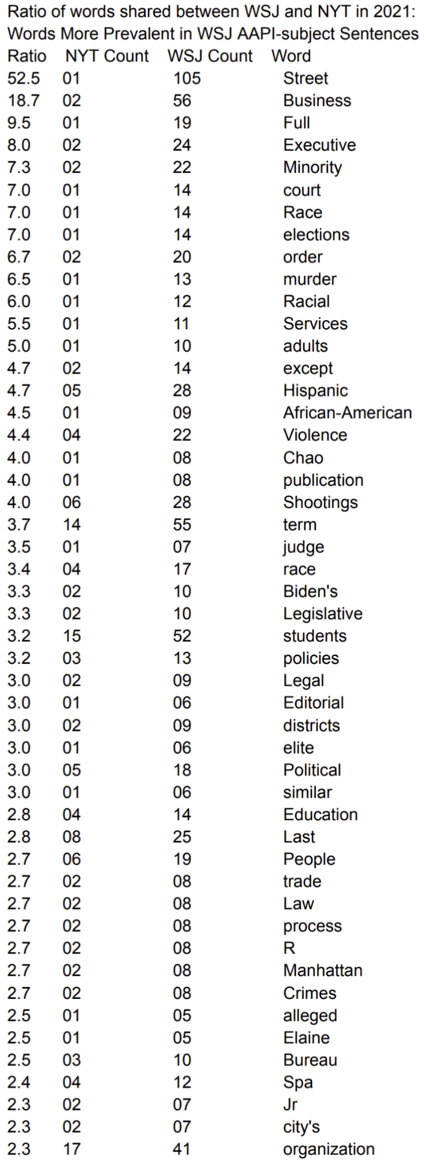
Table 5. Ratio of words shared between NYT and WSJ in 2021



Ratio of words shared between NYT and WSJ in 2021:
Words More Prevalent in NYT AAPI-subject Sentences

Ratio	NYT Count	WSJ Count	Word
54.0	108	01	Times
25.7	77	02	percent
24.0	72	02	3
18.5	37	01	reported
17.7	53	02	Section
16.4	131	07	2021
15.0	30	01	nearly
13.0	26	01	identity
12.0	24	01	politics
10.5	21	01	Kamala
10.4	52	04	voters
9.5	38	03	immigrants
9.0	27	02	10
9.0	45	04	End
9.0	27	02	Reserved
9.0	45	04	Document
9.0	18	01	face
9.0	45	04	States
9.0	27	02	reports
8.8	70	07	Stop
8.5	34	03	increase
8.4	67	07	like
8.2	33	03	Harris
8.2	74	08	Islanders
8.1	57	06	incidents
8.0	24	02	back
8.0	40	04	Copyright
8.0	16	01	much
8.0	16	01	either
7.9	63	07	2
7.8	47	05	among
7.7	23	02	Identity
7.5	15	01	decades
7.5	15	01	took
7.5	15	01	attention
7.5	15	01	physical
7.5	15	01	language
7.5	15	01	five
7.5	15	01	activists
7.5	15	01	surge
7.3	22	02	part
7.3	22	02	Park
7.3	22	02	news
7.0	14	01	February
7.0	21	02	fear
7.0	14	01	analysis
7.0	14	01	attacked
7.0	14	01	Women
7.0	14	01	election

Table 6. Ratio of words shared between WSJ and NYT in 2021:



Ratio of words shared between WSJ and NYT in 2021:
Words More Prevalent in WSJ AAPI-subject Sentences

Ratio	NYT Count	WSJ Count	Word
52.5	01	105	Street
18.7	02	56	Business
9.5	01	19	Full
8.0	02	24	Executive
7.3	02	22	Minority
7.0	01	14	court
7.0	01	14	Race
7.0	01	14	elections
6.7	02	20	order
6.5	01	13	murder
6.0	01	12	Racial
5.5	01	11	Services
5.0	01	10	adults
4.7	02	14	except
4.7	05	28	Hispanic
4.5	01	09	African-American
4.4	04	22	Violence
4.0	01	08	Chao
4.0	01	08	publication
4.0	06	28	Shootings
3.7	14	55	term
3.5	01	07	judge
3.4	04	17	race
3.3	02	10	Biden's
3.3	02	10	Legislative
3.2	15	52	students
3.2	03	13	policies
3.0	02	09	Legal
3.0	01	06	Editorial
3.0	02	09	districts
3.0	01	06	elite
3.0	05	18	Political
3.0	01	06	similar
2.8	04	14	Education
2.8	08	25	Last
2.7	06	19	People
2.7	02	08	trade
2.7	02	08	Law
2.7	02	08	process
2.7	02	08	R
2.7	02	08	Manhattan
2.7	02	08	Crimes
2.5	01	05	alleged
2.5	01	05	Elaine
2.5	03	10	Bureau
2.4	04	12	Spa
2.3	02	07	Jr
2.3	02	07	city's
2.3	17	41	organization

Next, when it comes to the ratio of words shared between WSJ and NYT in the post-pandemic period in 2021 (Table 6 above), words like "Street," "Business," "Executive," "Minority," "court," "order," "judge," and "Legal" are among the top words with very high ratios, suggesting a distinctive emphasis in WSJ's coverage of AAPI. These words suggest a strong business and legal focus in WSJ's coverage of AAPI subjects. Further, terms like "Race," "Racial”, "African-American," and "Hispanic" perhaps indicate a comparative focus on the racial and social aspect in the WSJ coverage of AAPI in relation to other racial groups in the US, such as African-Americans and Hispanic. In addition, higher word ratios of words like "Education," "People," "students," and "elite" suggest a focus on educational and cultural attributes of AAPI in WSJ, which provides more support for our third hypothesis regarding the perpetuation of ‘model minority’ stereotype about AAPI in conservative outlets. Lastly, terms like "Shootings," “murder”, "Violence," "Crimes," and "Spa" suggest a focus on crime and violence against AAPI in the post-pandemic period in WSJ coverage which was also seen in NYT coverage in 2021.

Topic Modeling Analysis Results:

Topic Modeling analysis provides additional support for our first hypothesis on the differences between liberal and conservative outlets, though the differences are once again more pronounced pre-pandemic. Again, in 2019 education-related words (“students”, “admissions”, “college”, etc.) dominated two of the four WSJ topics and zero of the four NYT topics. That year, only one of the four WSJ topics was dominated by politics-related words (“democratic”, “political”, “candidates”, etc.), while three of the four NYT articles were. This difference persisted in 2021, when one of the five WSJ topics was dominated by education-related words and none of the NYT were. Yet differences in 2021 were again less stark, as both outlets shifted some focus to attacks on AAPI and both had some focus on political issues and figures.

Table 3) Top 10 words per topic in 2019 NYT articles.

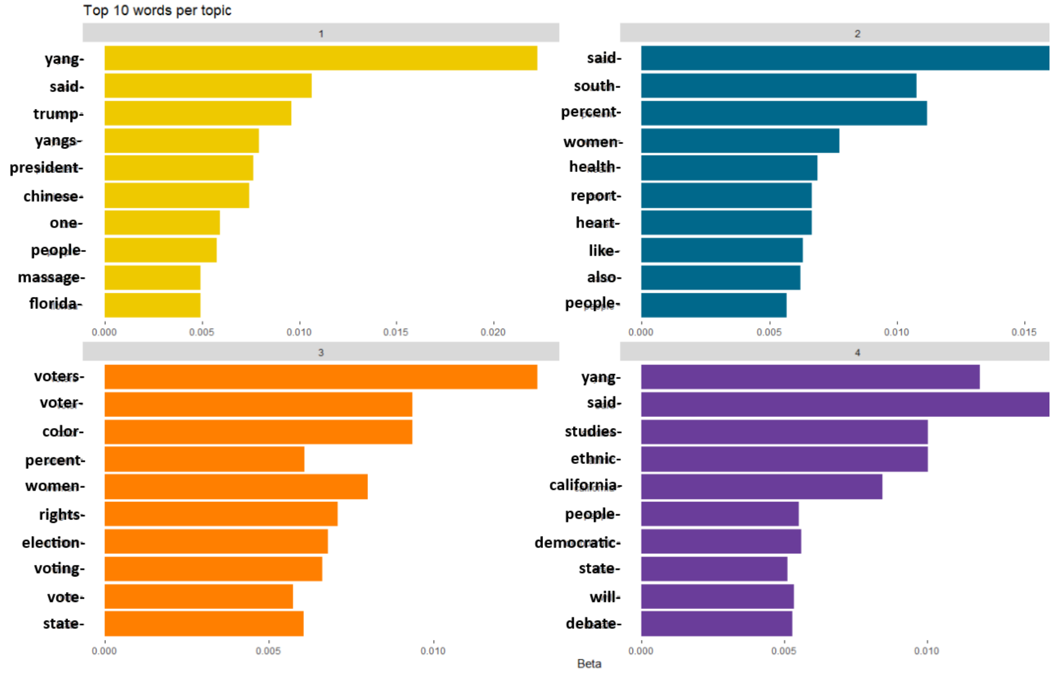
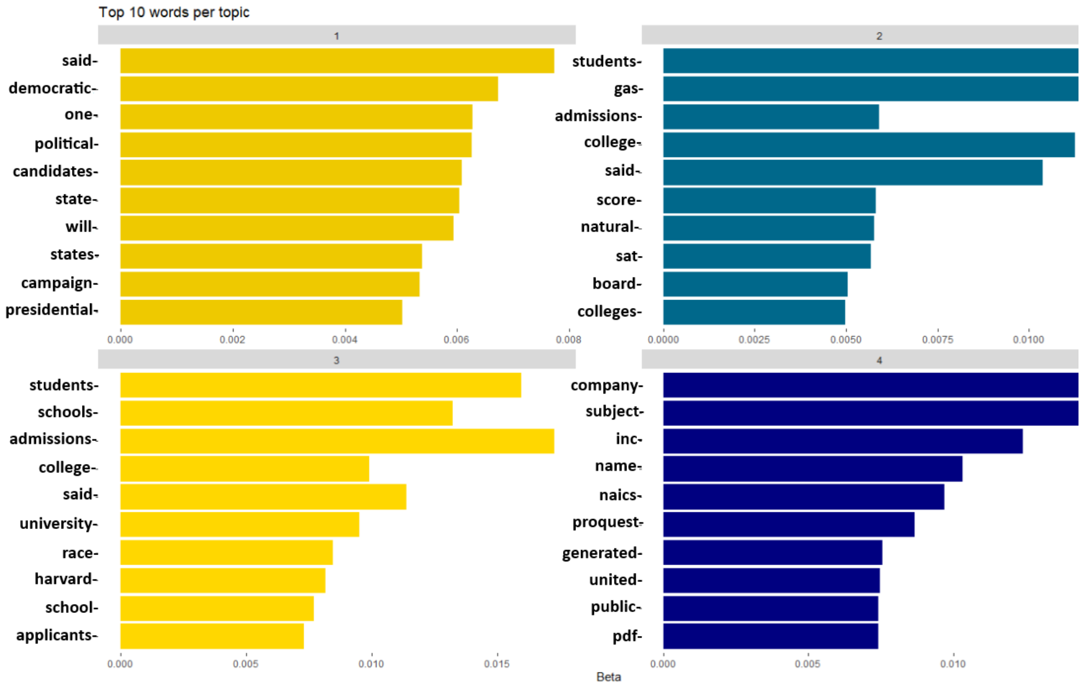


Table 4) Top 10 words per topic in 2019 WSJ articles.



Topic modeling analysis again provided support for our second hypothesis. The substance of coverage for both outlets differed in 2019 and 2021. And again, WSJ emphasis on educational words and topics provides some support for our third hypothesis regarding the perpetuation of stereotypes about AAPI in conservative outlets. As mentioned earlier, further qualitative study of the articles would be required to fully claim this as support for our hypothesis.

Table 5) Top 10 words per topic in 2021 NYT articles.

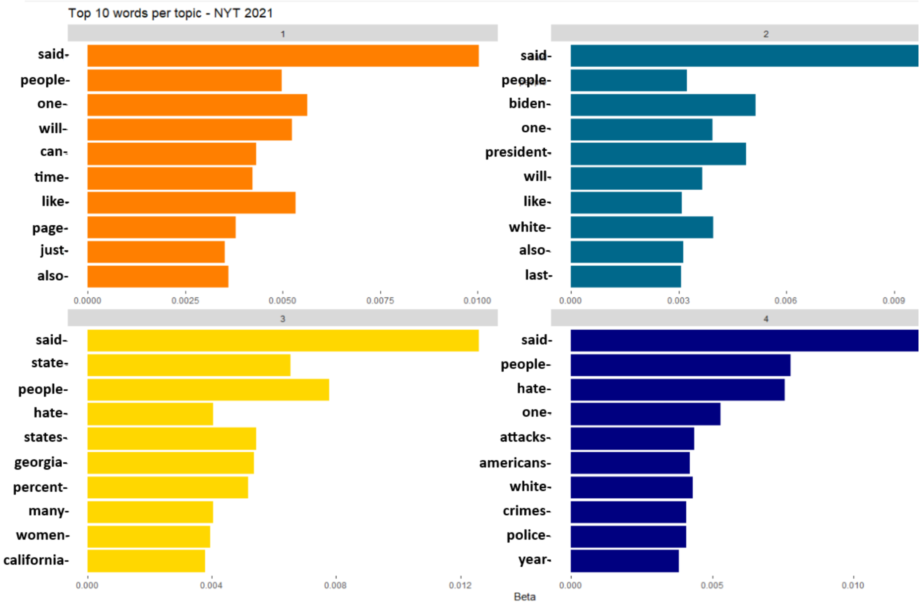
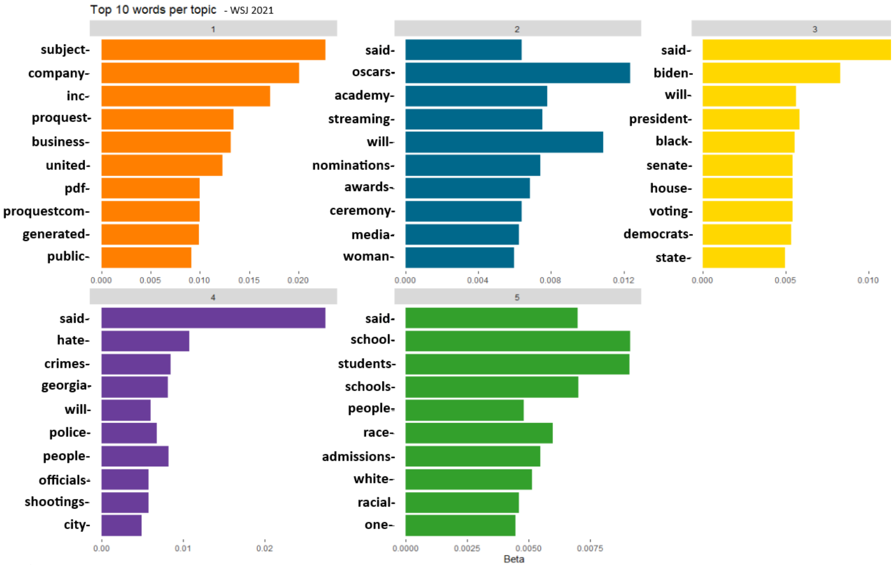
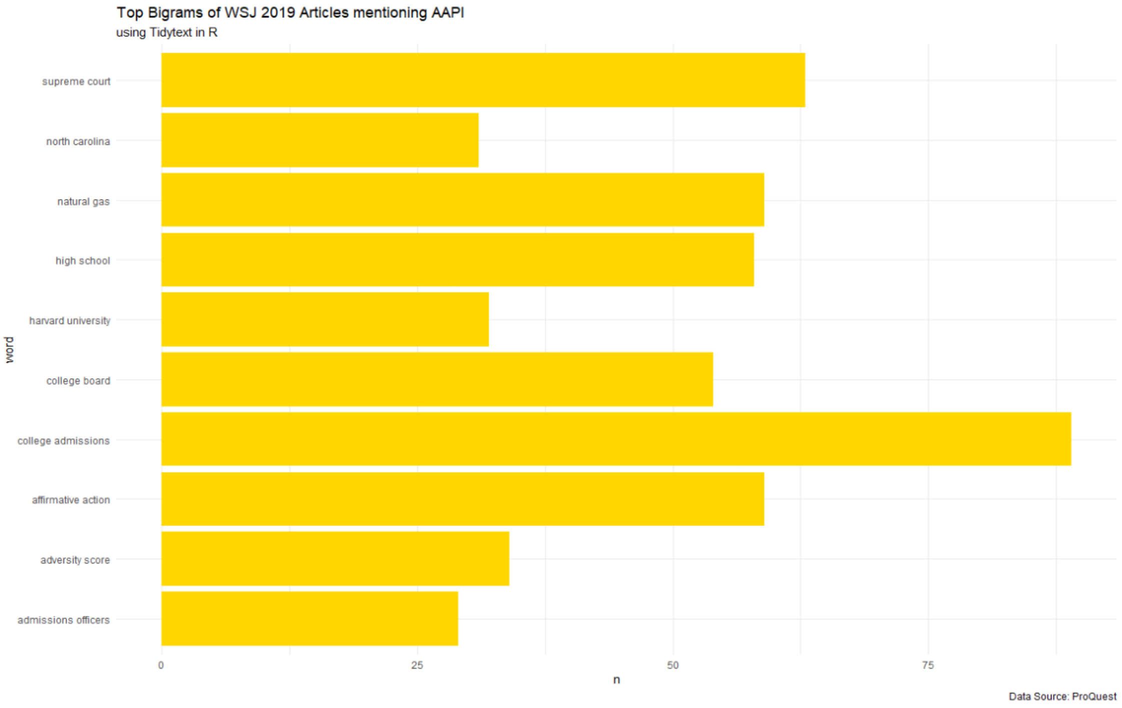


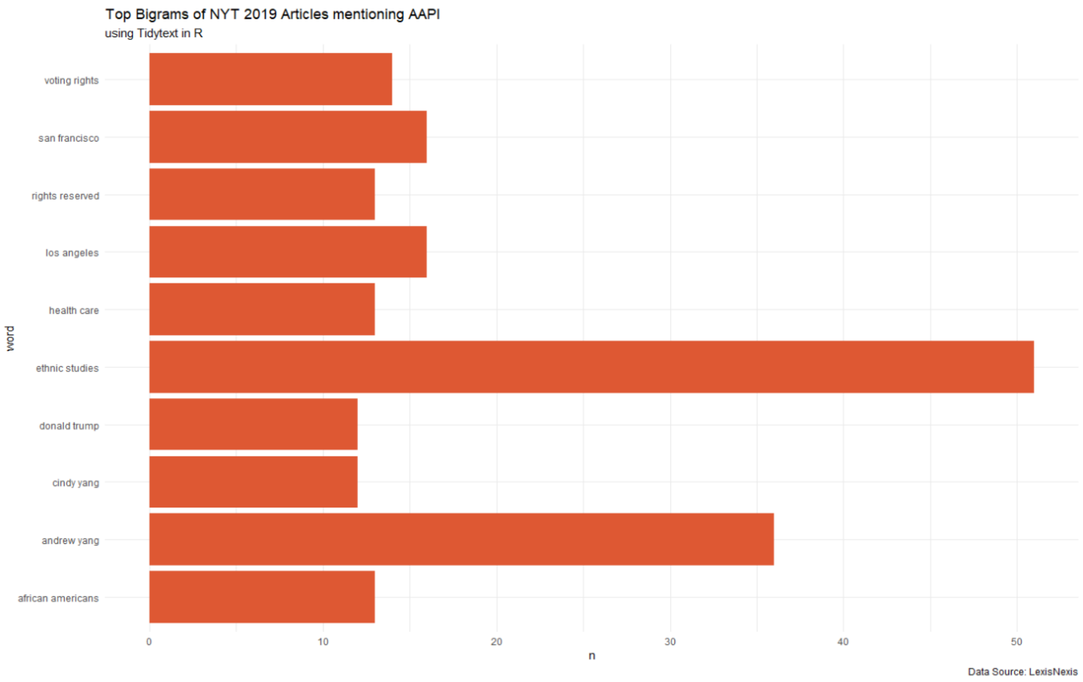
Table 6) Top 10 words per topic in 2021 WSJ articles.



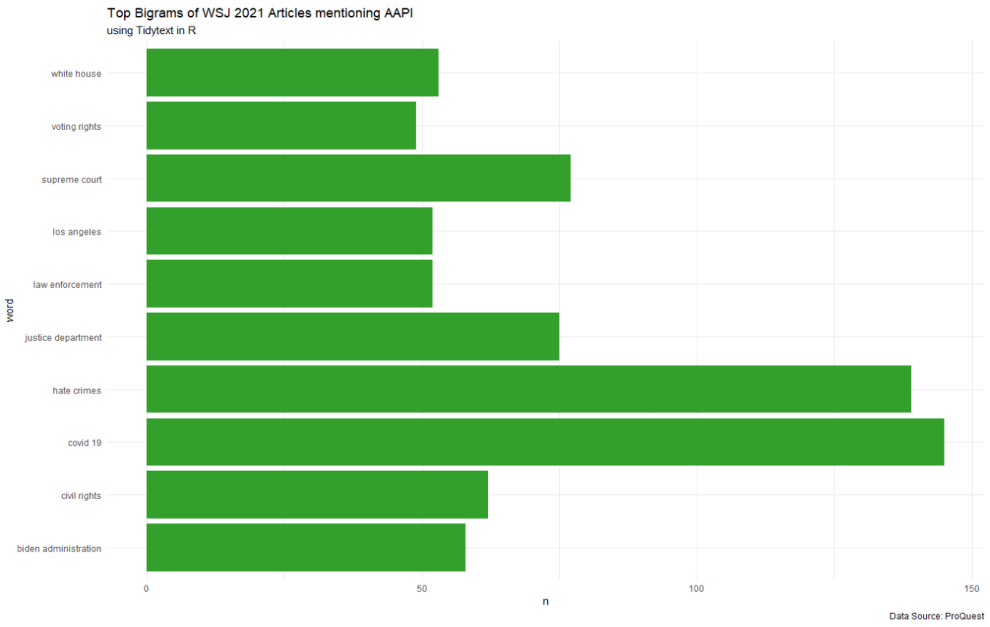
Bigram Analysis Results:

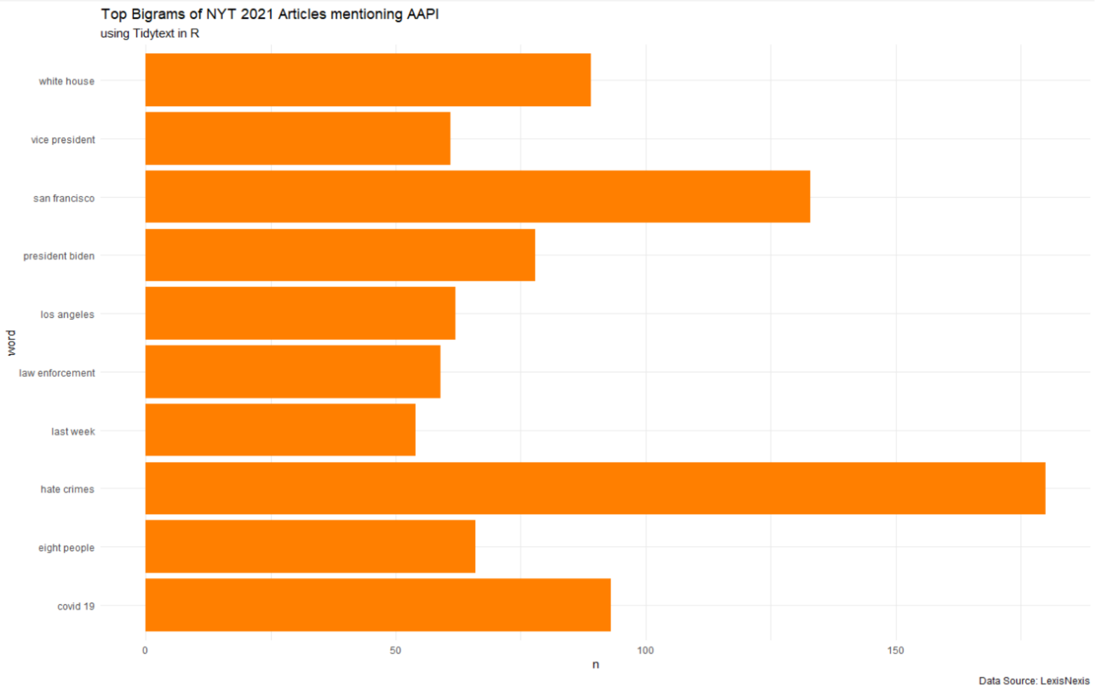
The bigram analysis provides additional support for our first (ideology) hypothesis, at least prior to the pandemic. The results again show stark substantive differences between coverage by WSJ and NYT in 2019. That year, WSJ had a high prevalence of bigrams related to education - 7 of the 10 most common bigrams in articles mentioning AAPI were education-related, including “high school”, “harvard university”, “college board”, “affirmative action”, “admissions officer,”, “adversity score“ and the most common bigram of all “college admissions”. In contrast, only one education-related bigram (“ethnic studies”) cracked the NYT top ten. The remaining nine bigrams touched on a variety of topics, from newsworthy individuals (Andrew Yang, Donald Trump and Cindy Yang), to legal and political issues (“voting rights”, “health care” and “rights reserved”) to locations with relatively large AAPI populations (San Francisco and Los Angeles) and to other groups (African Americans).





This difference vanishes in 2021, though, and the coverage in both outlets looks similar. For both NYT and WSJ, “hate crimes” and “COVID 19” are among the most common bigrams, as are political institutions and personalities (“President Biden”, “Biden Administration”, “Vice President”, “Supreme Court”, “White House”, “Justice Department”). This provides some support for our second hypothesis regarding a pre-pandemic and post-pandemic difference.





The bigram analysis also provides additional support for our third hypothesis about the perpetuation of stereotypes about AAPI in conservative outlets. As in previous analyses, the prevalence of education as a topic in 2019 WSJ articles suggests the “model minority” stereotype. But again, a more qualitative deep dive into reading the articles is needed to fully confirm that the prevalence of the bigrams corresponds with stereotyping. And here, the prevalence of these bigrams disappears after the pandemic and coverage by both outlets seems more similar in substantive terms in 2021. The differences that remain seem to be unrelated, or at least not obviously related, to stereotypes or other portrayals of AAPI as a group. For example, one difference is that the NYT retains a couple location-based bigrams (Los Angeles and San Francisco) which do not appear in the WSJ top ten and the WSJ has a couple rights-based bigrams (“voting rights” and “civil rights”) which do not appear in the NYT top ten.

Interpretation / Discussion:

The results of our analysis provide mixed support for our hypotheses. Here we discuss each hypothesis in turn:

Hypothesis 1: Conservative and liberal news outlets will portray AAPI differently

Our findings provide some support for this hypothesis. Bing sentiment analysis, bag of words analysis, and topic modeling analysis all found some differences between coverage of AAPI in WSJ and NYT. However, differences were more pronounced pre-pandemic than post-pandemic. This suggests that increased focus on discrimination and violence against AAPI post-pandemic may have overridden ideological differences, at least for these two outlets. For more extremely partisan news outlets, we may have seen differences in the post-pandemic era as well, as casual observation suggests that more liberal outlets were more concerned with defending AAPI and more conservative outlets may have been more vocal in blaming China for the virus itself, and by extension been less concerned about the welfare of AAPI.

Hypothesis 2: Portrayals of AAPI changed from pre-pandemic to post-pandemic

Our findings provide strong support for this hypothesis. All analyses found a significant difference in the tone of coverage of AAPI before and after the onset of the pandemic. For example, the proportion and prevalence of negative words - particularly those related to discrimination against and attacks on AAPI - was significantly higher in 2021 than it was in 2019. This suggests that the pandemic led to more attention to the negative treatment of AAPI.

Hypothesis 3: Conservative news outlets will be more likely to perpetuate stereotypes about AAPI than liberal news outlets.

Our findings provide some support for this hypothesis. The bag of words analysis, ratio of words analysis, and the bigram analysis found that WSJ had a greater emphasis on education-related terms than NYT. This suggests that WSJ may be more likely to perpetuate "model minority" stereotypes. However, the results of the Bing sentiment analysis did not find that the WSJ was more likely to use negative words about AAPI, even in the post-pandemic timeframe. Additionally, the differences in coverage between the two outlets were more pronounced pre-pandemic than post-pandemic. This suggests that the perpetuation of stereotypes about AAPI in the media may not be a perpetual part of WSJ coverage.

Overall, our findings suggest that the portrayal of AAPI in the national media is complex and nuanced. There are some differences between the coverage of AAPI in the conservative and liberal outlets that we chose, but these differences are not always consistent. Additionally, the portrayal of AAPI has changed significantly since the onset of the pandemic.

Strengths and Limitations of the Study:

Because of their proximity to the ideological center, the choice of the NYT and WSJ provides a scientifically conservative test of the hypothesis regarding a difference between politically conservative and liberal media outlets. This is both a strength and a weakness of our study. By providing a conservative test of the hypothesis, we have avoided "cherry-picking" data to fit our hypotheses. However, much of the actual American media universe is more extreme in partisanship and rhetoric than our chosen newspapers of record and analysis of such sources is likely to find more dramatic differences in coverage. Thus in terms of representativeness, the study could be strengthened by using a larger sample of articles from a wider variety of sources.

Another strength of the study is the breadth of our coverage. Using R and Python scripts allows us to gather data from a much larger sample than would reading each article individually. This provides a more accurate view of the overall coverage of the topic provided by each source than would choosing to analyze a small sample from each year. Still, what we gain in breadth, we lose in depth. A deeper, more qualitative study in which we read a sample of articles from each source and year would help definitively confirm the presence of stereotyping in reporting.

Lastly, a strength of our analysis is in its focus on the news media, which is the main source by which people in the United States learn about and interpret the events of the day. Yet it is not the only information source that shapes the public's view of groups of people like AAPI.

The study could be further extended by examining the sentiment of coverage of AAPIs in other types of media, such as television and film.

Conclusion:

The portrayal of Asian American Pacific Islanders in news outlets matters for how people in the US react to them. Our study demonstrates noticeable differences in the portrayal of AAPI in conservative and liberal news outlets. For example, NYT tended to emphasize politics-related terms, while WSJ emphasized education-related terms. This suggests liberal outlets may be more likely to portray AAPI as political actors, while conservative outlets may be more likely to engage with AAPI through model minority stereotypes.

We also present evidence that coverage of AAPI differs before and after the COVID-19 pandemic. One notable difference was a significant increase in the proportion and prevalence of negative words in coverage after the pandemic, suggesting that discrimination and violence against AAPI received more attention. This is consistent with other research that has found an increase in anti-Asian sentiment during the pandemic.

Our findings have several important implications. First, they suggest that news outlet partisanship can influence how AAPI communities are perceived. This is because news outlets may frame AAPI issues in ways that align with their partisan agendas. Second, our findings suggest that the COVID-19 pandemic led to a substantive change in the coverage of AAPI. This is likely due to the fact that the pandemic was associated with a rise in anti-Asian sentiment in the United States. While our study adds in important ways to the literature, it is also limited in important ways, and as such further research to better understand the complex relationship between media portrayals and AAPI communities is necessary.

Work Cited:

Ad Fontes Media. (2023). Media Bias Chart® 11.0 - Static Version . Retrieved November 26, 2023, from https://adfontesmedia.com/static-mbc/

Baum, M. A., & Peters, E. (2013). The influence of news media on public opinion: A theoretical and empirical overview. In Sage handbook of political communication (pp. 3-32). SAGE Publications.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4-5), 993-1022.

Farrell, H. (2007). Framing news: Media, politics and the construction of public opinion. Pluto Press.

Iyengar, S., & Kinder, D. R. (1987). News that matters: Television and American opinion. University of Chicago Press.

Jang, S. Y., Wong, Y. J., & Seah, K. (2021). COVID-19 pandemic and the negative media portrayal of Asian Americans and Pacific Islanders (AAPI). PLOS One, 16(11), e0260095.

Napoli, P. (2011). Media studies: An introduction. Routledge.

SANTIA, M., ODEN, A., SEON-WOO KIM, PINGREE, R. J., WYERS, J., & BRYANOV, K. (2022). The Other Side of the Pandemic: Effects of Racialized News Coverage on Attitudes Toward Asians and Immigrants. *International Journal of Communication* (19328036), 16, 5717–5739.

Shoemaker, P. J., & Vos, T. (2008). Gatekeeping and the news process. Routledge.

Sue, D. W., Lin, A. I., Torino, G., Capodilupo, C. M., & Rivera, D. P. (2007). Racial microaggressions toward Asian Americans: Explicit and subtle forms and their psychological impact. Journal of Counseling Psychology, 54(4), 474-487.

Tsai J, Phua J, Pan S, Yang C. (2020. Intergroup Contact, COVID-19 News Consumption, and the Moderating Role of Digital Media Trust on Prejudice Toward Asians in the United States: Cross-Sectional Study. *J Med Internet Res 2020;22*(9):e22767 doi: 10.2196/22767

U.S. Census Bureau. (2022). Population estimates by race and ethnicity. https://www.census.gov/quickfacts/fact/table/US/PST045221