Racial Bias in Media

Client: ***NAACP Boston Branch***

Team Members

Neeza Singh - BS Data Science, Class of 2025 | nsingh03@bu.edu
Ishan Ranjan - MS Data Science, Class of 2024 | iranjan@bu.edu
Hannah Choe - BS Data Science, Class of 2025 | hanchoe@bu.edu
Abdul Rafay - BS Data Science, Class of 2026 | rafaya@bu.edu
Derinell Rojas - BS Computer Science, Class of 2025 | droja@bu.edu
Zachary Gou - BS Computer Science, Class of 2024 | zgou@bu.edu
John Markowicz - BS Data Science, Class of 2024 | jcmark@bu.edu

BOSTON UNIVERSITY

INTRODUCTION & PROBLEM STATEMENT

In an era increasingly attuned to racial disparities and the progression of social justice initiatives, the media's portrayal of underrepresented communities wields significant influence over public perceptions and societal narratives. Our project undertakes an analysis of The Boston Globe's coverage of the Black community in Greater Boston, particularly articles published from 2010 through 2019. This period, critical for its social dynamics, sets the stage for our inquiry into the intersection of media representation and racial discourse.

Our analysis aims to address key questions:

1. How does the portrayal of sentiment in media coverage vary by location within the Greater Boston area, and what relationship does this bear to the racial and ethnic demographics predominant in those locales?
2. In what manner is the distribution of media articles skewed by race and ethnicity, and how does coverage volume across communities compare?
3. In the realm of editorial content, how do the narrative tone and focus differentiate from other article types when they engage with topics concerning diverse racial communities and geographic sectors?

ABOUT OUR DATASET: CONSTRAINTS & STRUCTURE

The dataset for our analysis is drawn from LexisNexis, capturing a decade of The Boston Globe’s reporting from the years 2010-2019. The exclusion of the year 2020 was a methodological choice, necessitated by an incomplete dataset for that year, which could potentially skew the integrity of our analysis.

Each article within our dataset is meticulously cataloged with several key attributes, maintaining consistency in data structure across the decade. These attributes include 'position_section' and 'position_subsection', which categorize the article within the newspaper's layout, providing a foundational context for its thematic placement. Additionally, every article is accompanied by its headline and body text, offering direct insights into the article's content and thematic orientation. The 'lede', a succinct introduction to each article, and a corresponding collection of keywords enrich the dataset further by offering highlighting dominant themes.

Navigating through the dataset's architecture presented distinctive challenges. The 'position_section' column effectively guided the classification of articles into recognizable segments of the newspaper such as 'Editorial Opinion,' 'Sports,' and 'Business.' In contrast, the 'position_subsection' column did not yield the expected granularity, prompting us to seek alternative methods for discerning the articles' primary topics.

The task of cataloging each article by its specific location was significantly challenged by the extensive corpus of 177,000 articles within the constrained time frame of our study. To navigate this logistical challenge, we developed a streamlined subset of the data, to increase the computational speed of our demographic-based analysis. We also noted that a considerable portion of the articles broadly mentioned 'Boston' or 'Massachusetts' or did not meaningfully specify a location. This observation necessitated a focused geospatial analysis, yet only a fraction of the articles (~1000) could be definitively associated with precise geographic locations and demographic information.

Overall, the dataset lays a robust foundation for an in-depth examination of media representations' geographical and racial dynamics. This analysis is designed to uncover the intricate patterns in journalistic practices over the past decade, thereby enriching our comprehension of the nuanced portrayals of racial and ethnic communities within the media sphere.

METHODS

Data Cleaning

To ensure uniform data processing, we combined all datasets into a singular .csv file. A specialized data cleaning function was developed to standardize the text within the “hl1”, “hl2”, and “body” columns by transforming it to lowercase and eliminating extraneous whitespaces. We excised the "Unnamed", "content-id", and "word_count" columns due to inaccuracies in the provided word counts. In their place, we introduced a new column, "actual_word_count", reflecting the precise word count of the body text. Furthermore, we employed regular expressions (Regex) for the refinement of the body text and utilized the NLTK Python package to filter out stopwords, thereby enhancing the dataset's cleanliness and relevance for our analysis.

Geocoding

Geocoding is composed of multiple steps. We break down the pipeline into two parts. Part one is the semantic interpretation of the articles and pulling out locations. The second part consists of testing each possible location for an article if it is within Boston or not, starting from the most specific to the least specific mention of location. Part one consists of two models. The article is first passed through a Large Language Model (LLM) to cover a semantic interpretation with the chosen architecture being Meta's Llama 7 Billion Parameter model. From here, this guarantees that outputs of text in the pipeline are well formatted making the pipeline robust. We then use Span_Marker's Transformer Named Entity Recognition (NER) model to pull locations out of the LLM response leaving a collection of possible locations with varying specificity for each article. This step will be discussed in the next section in **Census API Mapping.** Although our pipeline discussed is so far robust, the problem is hinted at in computation power and time. Given our dataset of approximately 177,000 articles, it would *take a few days* to process it all on modern computers.

Census API Mapping

As mentioned above, each article is composed of a list of locations. A given example is below for an article:

"[(Boston, 'GPE'), (Massachusetts, 'GPE'), (1, 'CARDINAL'), (Boston, 'GPE'), (Massachusetts, 'GPE'), (2, 'CARDINAL'), (Boston, 'GPE'), (South End, 'LOC')]"

With each label of GPE, CARDINAL, LOC, etc. following tags of the `IOB`, `IOB2`, `BIOES`, `BILOU` formats. From here, we filter and distill the locations with several conditions on the labels. We would like to cast out uninformative labels. Additionally, we filter based on specificity, being the most specific taking priority. This guarantees we label each article with one location. Such a list after filtering would result in:

"[('South End', 'LOC')]"

We then take such a location and give it to the MapBox API to obtain coordinates in longitude and latitude. The given coordinates are then given to the U.S Census API to obtain the corresponding census tract. With the census tract, information such as majority demographics, county, neighborhood, and more can be derived.

Sentiment Analysis

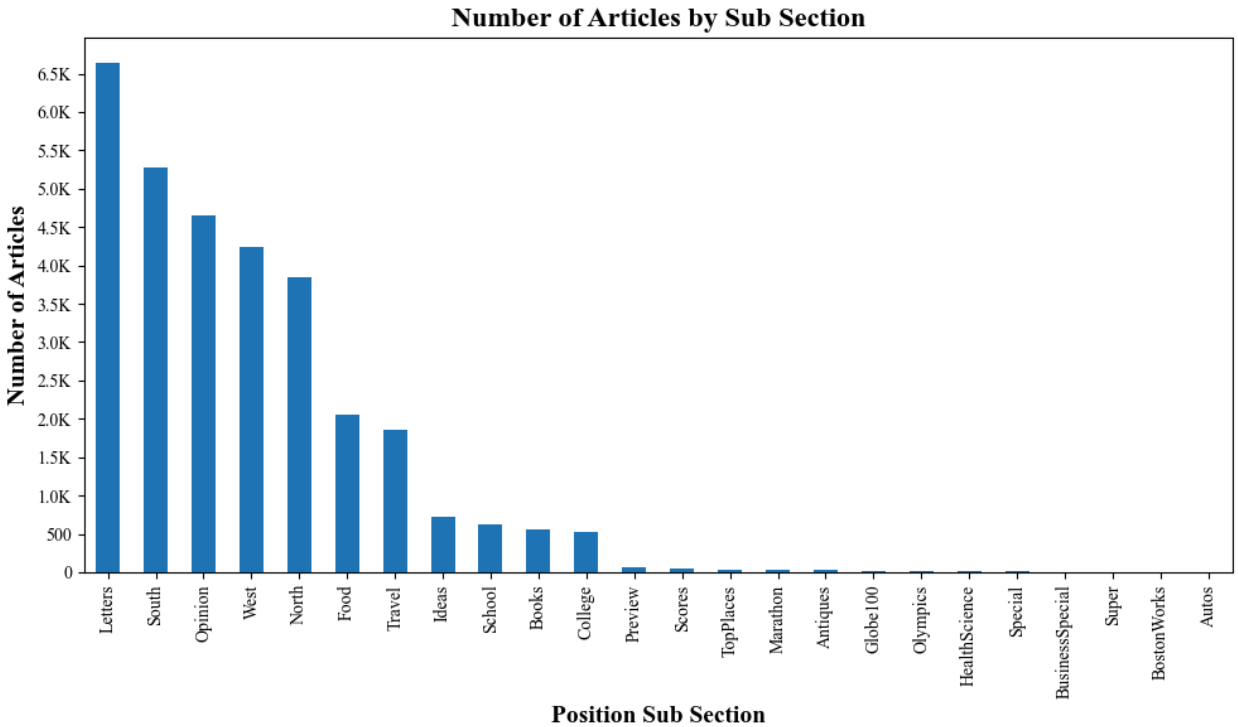
For sentiment analysis, we utilized the NLTK module called Valence Aware Dictionary for Sentiment Reasoning (VADER). This module provides the Sentiment Intensity Analyzer which was utilized to compute the opinion metric of the article's texts. The model is trained on a pre-made corpus of words and their connotations, allowing us to parse through the body, headings, and keywords for each article.

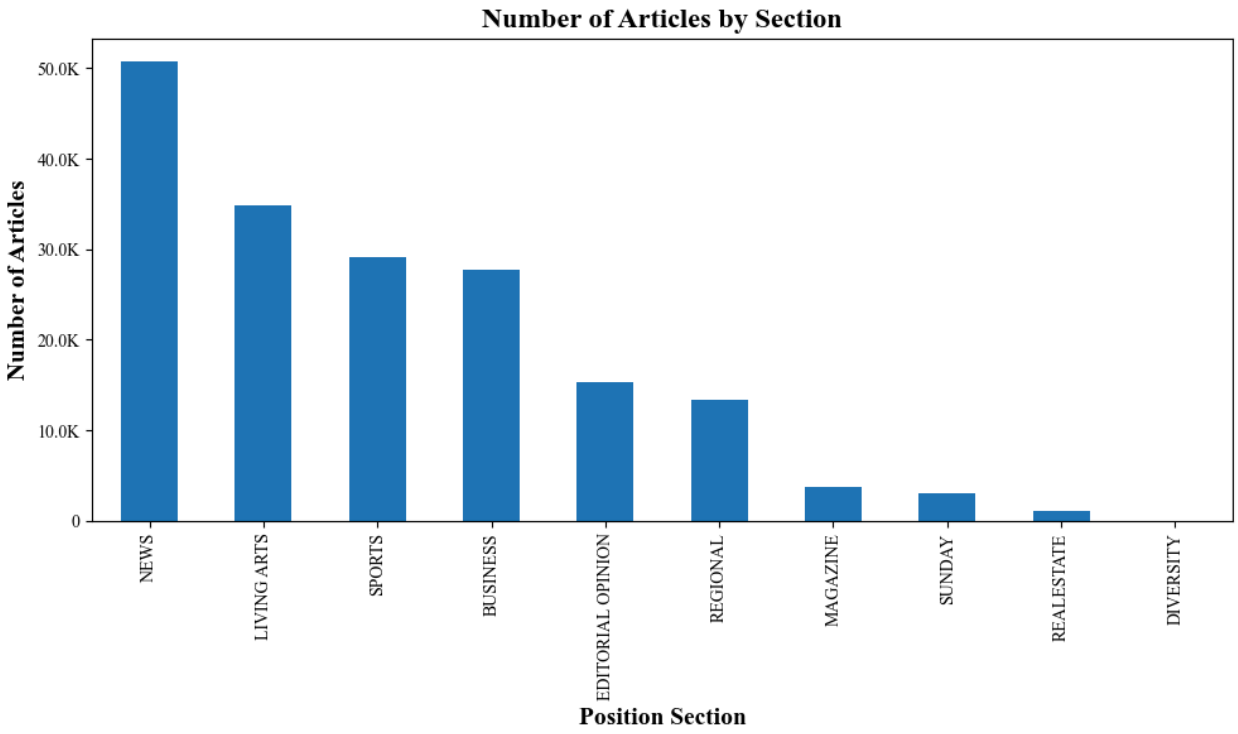
RESULTS AND INSIGHTS

Exploratory Distributions

Distribution of Articles by Position Section & Subsection

The visualization offers a preliminary look into the distribution of all ~177,000 articles by section within The Boston Globe, serving as a contextual foundation for our dataset. It reveals a marked prominence of the "News" section, affirming its status as a central hub for reporting, followed by the "Living Arts" and "Sports" sections, which also contribute significantly to the overall content. Sections such as "Editorial Opinion", "Regional", and "Magazine", while less represented, offer a glimpse into the diverse range of topics covered by the publication.



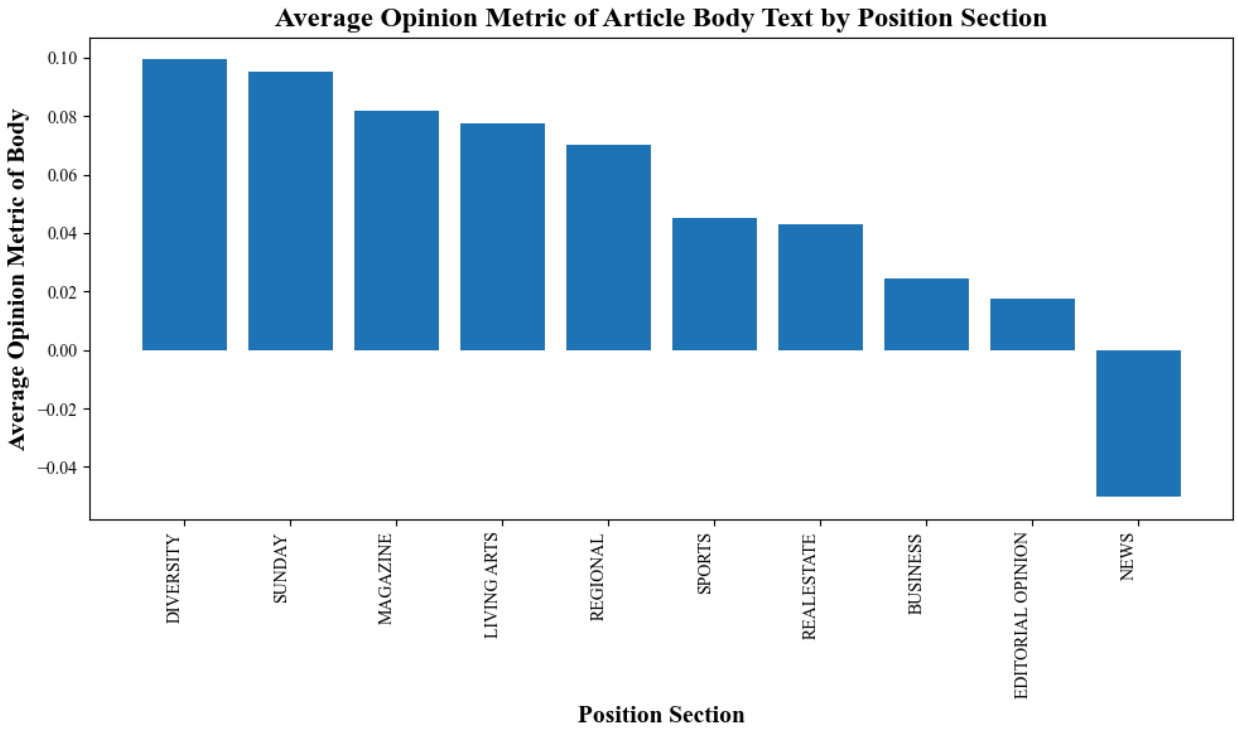


For this visualization, the sub-sections without bars represent duplicate categories and/or areas with fewer than ten articles, indicating limited coverage in our dataset.

Opinion Metric & Sentiment

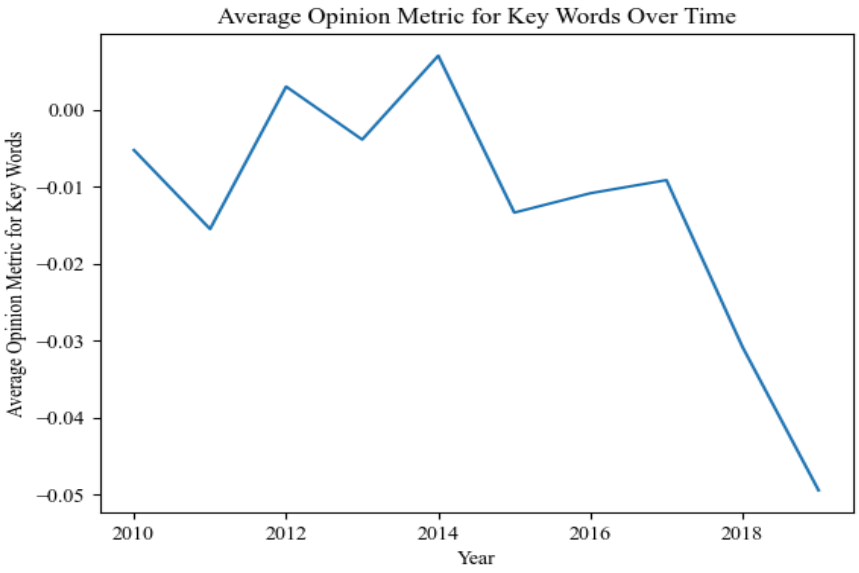
Average Opinion Metric by Position Section

The bar graph below offers a comparative view of the average opinion metrics by position section within The Boston Globe. It reveals that sections like 'Diversity' tend toward a more positive metric, whereas 'News' is characterized by a lower, more negative metric, suggesting a more critical tone in the content. Notably, the 'Sunday' section's metric, while higher, should be interpreted with caution due to its topic ambiguity. Such preliminary findings lay the groundwork for a nuanced understanding of the publication’s editorial voice and how it might vary across different sections.



Average Opinion Metric Over Time

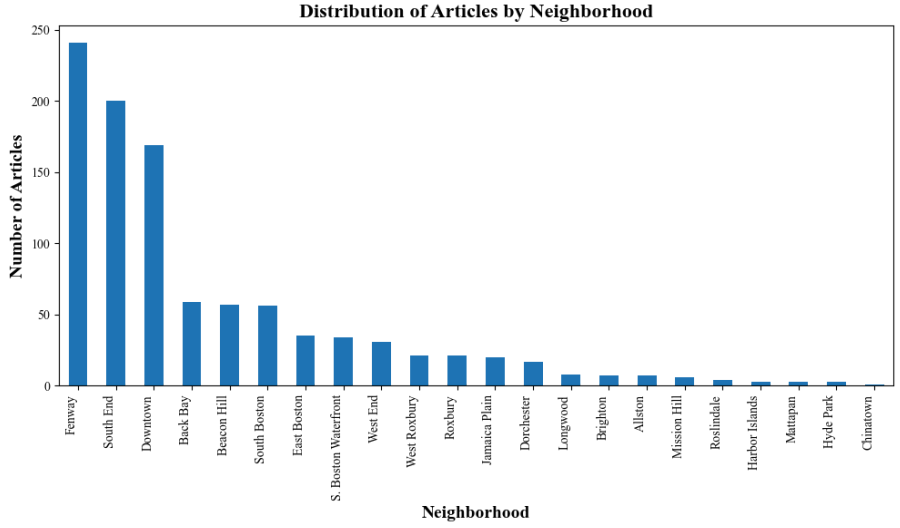
The graph presents a temporal analysis of sentiment scores derived from keywords in The Boston Globe's articles. Sentiment scores range from positive to negative, with scores above zero indicating a more positive tone and scores below zero denoting a more negative or disapproving stance. Scores close to zero are indicative of a neutral or balanced tone. Notably, there has been a pronounced shift toward lower sentiment scores since 2016, reflecting a possible intensification of criticality in the narrative around these topics. This data may mirror the broader societal and political climate, hinting at the influence of external events on journalistic sentiment.



Location and Race-based Analysis

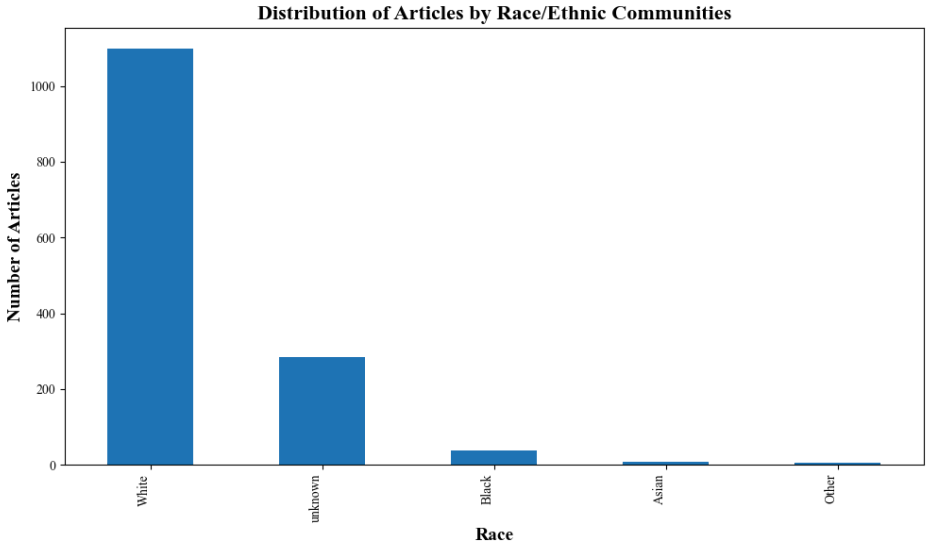
Distribution of Articles by Location

The bar graph delineates the distribution of articles across Boston's neighborhoods, highlighting the concentration of media attention in specific areas. The preeminence of neighborhoods like Fenway and South End suggests a particular editorial focus, while areas such as Chinatown and Hyde Park exhibit considerably less coverage. This visualization is crucial at the foundational level, as it sets the stage for probing into the nuances of racial bias in media. The geographic emphasis revealed by the data may correlate with demographic profiles, allowing us to question whether disparities in article volume reflect broader patterns of representation or oversight in the context of racial and community narratives within the city.



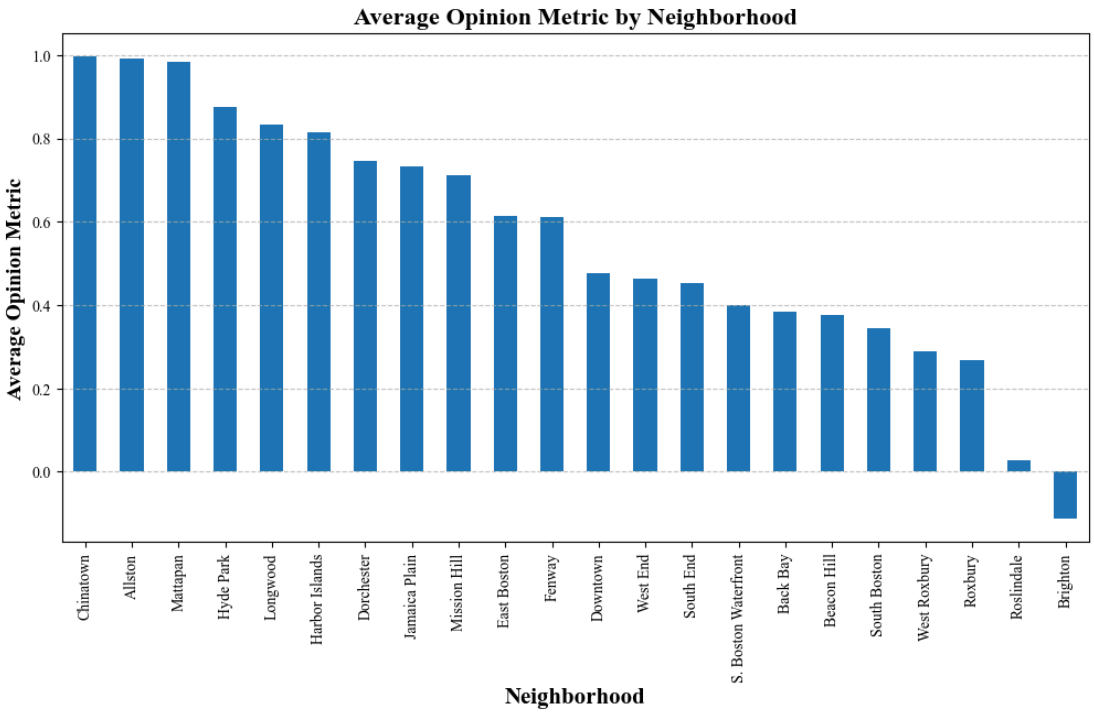
Distribution of Articles by Race

The bar graph illustrates the distribution of The Boston Globe's articles by race/ethnic communities, with a pronounced skew toward the White demographic. The volume of articles referencing Black, Asian, and other racial groups is noticeably lower. This pattern not only suggests potential disparities in media representation but also may intersect with the geographic distribution of Boston's neighborhoods shown in the previous graph. The concentration of coverage could reflect the demographic makeup of these neighborhoods or indicate a narrative bias. Together, these visualizations underscore the importance of spatial and racial context in media coverage



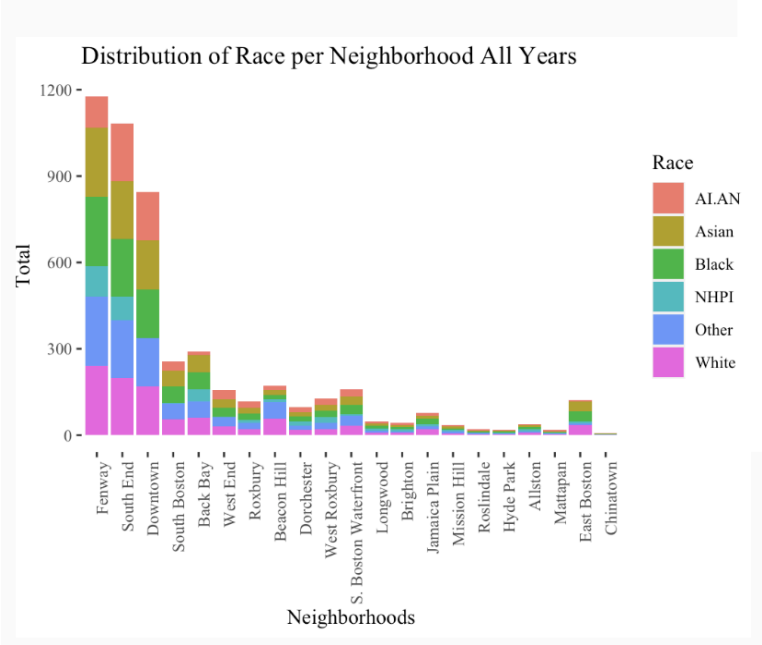
Average Opinion Metric of Articles by Neighborhoods

In order to further research the coverage of certain neighborhoods in Boston, we looked at how articles about specific neighborhoods stood on the opinion metric. As shown by the visualization below, there is a high variance in which neighborhoods have a high opinion metric and a lower one. Given the next visualization of the distribution of race by neighborhood, this seems to corroborate the finding of a lack of coverage of black communities and neighborhoods, as majority-white neighborhoods lie all over the distribution. However, the presence of neighborhoods such as Roxbury near the bottom of the opinion metric rankings may suggest racial bias in that such minority neighborhoods may be covered more negatively than others.



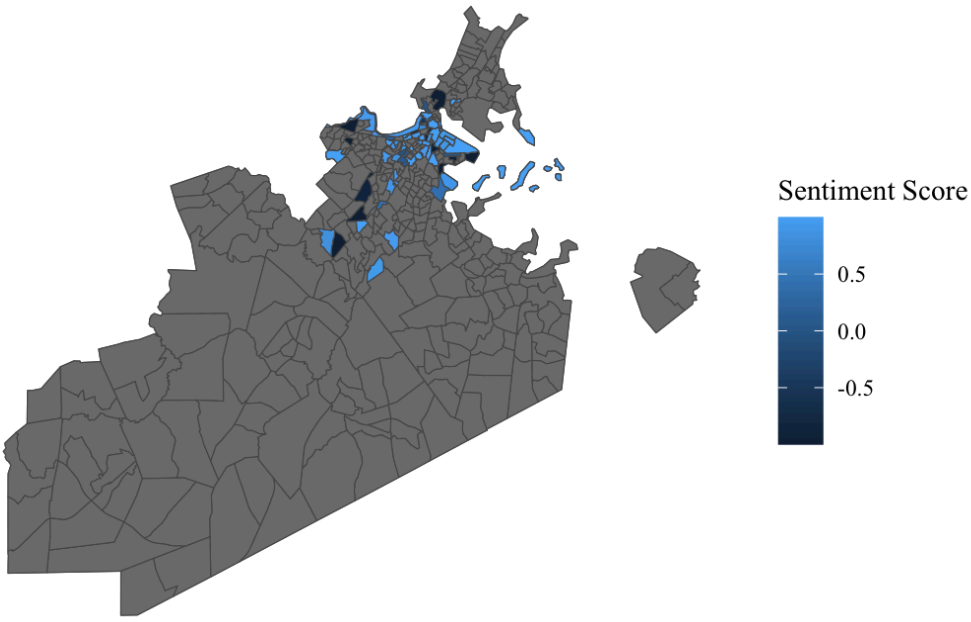
Distribution of Racial Groups by Location

To double-check our findings, we took a look at the distribution of racial groups by neighborhood in Boston, which is visualized below:



Opinion Metrics across Counties in Greater Boston

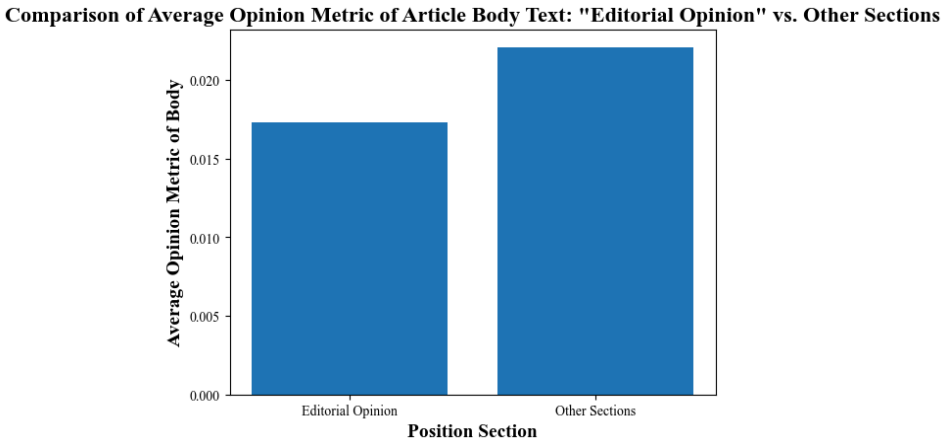
As shown by the following visualization of sentiment scores on a map of Boston, we can see that neighborhoods such as South Boston and Everett have more negative opinion metrics/sentiment scores. This is also true for southern Boston neighborhoods such as Jamaica Plain.



Opinion Editorial Analysis

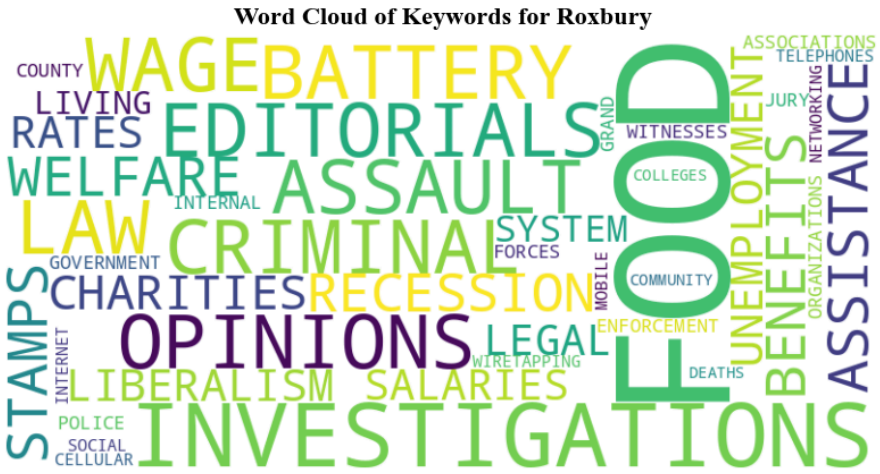
Average Opinion Metric of Article Body for Op-Eds vs Other Content

To further understand bias in specific sections of the newspaper, we took a look at editorial opinions (Op-Eds). This is because these editorials tend to be more free-form and are potentially less restrained from journalistic standards of avoiding bias in reporting. Looking at the average opinion metric for op-eds vs. other sections as shown below, we can see that the general values for the opinion metric/sentiment of Op-Eds seem to be lower than that of the other sections, while still positive, seeming to suggest more neutrality in these sections.

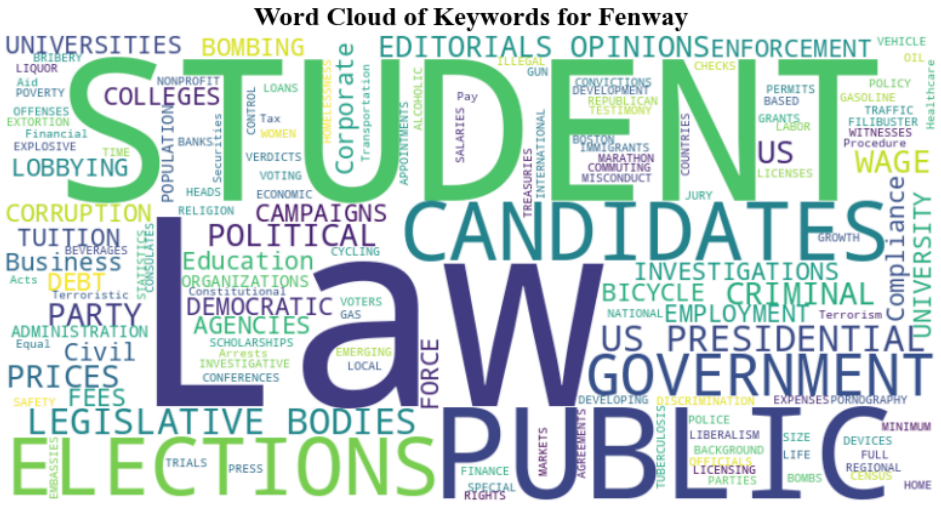


Below are visualizations of the most common words in article bodies about the neighborhoods of Roxbury and Fenway

Most Commonly Used Words in OpEds about Roxbury



Most Commonly Used Words in OpEds about Fenway



CONCLUSION

Overall, based on the articles that we were able to process and geolocate, we were able to draw the following key insights from the data:

(1.) More positive leaning in Fenway and Downtown Boston (majority white) and sparse spots in other neighborhoods.

(2.) Despite the small sample size, we find that the majority of articles are written in white-majority neighborhoods compared to other races and ethnicities. Although the "NEWS" section is the most frequent major topic, it portrays a very general category that may not lead to impactful insights. Rather, observing the following major categories of "LIVING ARTS" and "SPORTS", we see that such articles fall majority to white-majority neighborhoods.

(3.) Furthermore, from the perspective of narrative tone and focus based on article types, we see that articles that are generally opinionated such as op-eds have negligible sentiment between other article types. The figures suggest that what is being reported has a stronger correlation to geographic location rather than sentiment alone. As discussed before, such can be seen with keywords from articles geolocated in Roxbury compared to articles geolocated in Fenway.

Some challenges that we faced while completing this analysis included a lack of computing power to adequately run large deep learning LLM models and locate all of the locations mentioned in the articles, as well as a limited scope of time to complete the project. However, we are confident that with computing power and time, that methods

such as geolocation and LLM-based location extraction, combined with sentiment analysis and key word overviews, can provide valuable insights that can help inform decisions that will hopefully make Boston a more equitable city.

DELEGATION OF TASKS

Neeza Singh - Team Lead, Data Cleaning, Sentiment Analysis, Visualizations
Ishan Ranjan - Scrum Master, Data Cleaning, Sentiment Analysis, Visualizations
Hannah Choe - Data Cleaning, Sentiment Analysis, and Visualizations
Abdul Rafay - Data Cleaning, Geolocation, and Census Map API
Derinell Rojas - Data Cleaning, Geolocation, and Census Map API
Zachary Gou - Data Cleaning, Geolocation, and Census Map API
John Markowicz - Data Cleaning,