

NAACP Media Research Project
CS 506 - Spark! Project
Nada Shalash | Shubhangi Jain

I. Summary and Problem Statement

The National Association for the Advancement of Colored People (NAACP) is the country's largest and most widely-recognized civil rights organization, founded in 1909 with the goal of addressing racial disparities in the US and ensuring equality of political, educational, and economic rights. Our client was the Boston chapter of the NAACP.



The Boston Globe

For this project, the client was interested in analyzing racial bias and race-based discrimination in media coverage by media outlets in Boston. The intention was to use the results of this project and present them at the NAACP annual meeting which was supposed to take place in Boston later this year.

The goal of this project was to continue building on the work that was conducted by last semester's teams to gain a better understanding of media coverage of Black and African-American communities in the Boston area, using articles from the Boston Globe from 2014 to 2018 and focusing on three main areas:

1. Analyzing coverage of neighborhoods and sub-neighborhoods with a higher percentage of Black residents relative to coverage of other neighborhoods using sentiment analysis and noting any differences or disparities in this coverage
2. Analyzing coverage of Black homicides by comparing the volume of coverage relative to white homicides and conducting sentiment analysis
3. Examining articles that explicitly mention race, using keywords for Black such as "African-American" or "Haitian-American" and determining the common topics and themes of those articles

II. Methodology and Algorithms

- **Coverage of neighborhoods and sub-neighborhoods:**

Define Neighborhoods and Sub-Neighborhoods:

Suffolk county, which encompasses the greater Boston area, has 23 official neighborhoods. Last semester's team conducted their analysis using these

neighborhoods. However, this semester, we were asked to also include an analysis of sub-neighborhoods, since some neighborhoods might have their minority population concentrated within a certain part of the neighborhood, and taking this into consideration would make the analysis more robust.

In order to define the neighborhoods and sub-neighborhoods, we downloaded US Census Tract data as well as the demographic information for all of the census tracts in Suffolk County. Our team collected the first half of the census tract demographic data and the second team collected the second half. This is what the spreadsheet looked like:

| Census Tract Number | Percent Male | Percent Female | Total Population | One Race | | | | | | Total Population | White and Black or African American |
|---------------------|--------------|----------------|------------------|----------|---------------------------|-----------------------------------|-------|--|-----------------|------------------|-------------------------------------|
| | | | | White | Black or African American | American Indian and Alaska Native | Asian | Native Hawaiian and Other Pacific Islander | Some other race | | |
| Census Tract 1 | 50.30% | 49.70% | 4,668 | 2,693 | 274 | 0 | 1,044 | 0 | 599 | | 9 |
| Census Tract 2.01 | 46% | 54% | 3,857 | 3,120 | 198 | 0 | 393 | 0 | 73 | | 33 |
| Census Tract 2.02 | 50.10% | 49.90% | 3,982 | 2,769 | 336 | 0 | 470 | 0 | 372 | 35 | 5 |
| Census Tract 3.01 | 58.20% | 41.80% | 2,905 | 2,069 | 170 | 0 | 434 | 0 | 223 | 9 | 0 |
| Census Tract 3.02 | 48.90% | 51.10% | 3,385 | 2,258 | 387 | 0 | 533 | 61 | 107 | 39 | 1 |
| Census Tract 4.01 | 43.40% | 56.60% | 5,112 | 3,626 | 66 | 0 | 1,238 | 0 | 31 | 151 | 80 |
| Census Tract 4.02 | 49.90% | 50.10% | 3,431 | 2,575 | 93 | 0 | 556 | 0 | 10 | 197 | 37 |
| Census Tract 5.02 | 45.20% | 54.80% | 5,881 | 4,921 | 126 | 11 | 623 | 0 | 45 | 155 | 22 |
| Census Tract 5.03 | 51.60% | 48.40% | 2,123 | 1,704 | 40 | 15 | 153 | 0 | 39 | 172 | 0 |
| Census Tract 5.04 | 41.60% | 58.40% | 4,449 | 3,491 | 106 | 78 | 570 | 0 | 29 | 175 | 79 |
| Census Tract 6.01 | 52.30% | 47.70% | 3,800 | 2,641 | 123 | 14 | 737 | 0 | 190 | 95 | 37 |
| Census Tract 6.02 | 39.30% | 60.70% | 3,896 | 2,089 | 425 | 11 | 755 | 0 | 524 | 92 | 11 |
| Census Tract 7.01 | 49.20% | 50.80% | 4,296 | 2,884 | 350 | 64 | 832 | 39 | 21 | 106 | 0 |
| Census Tract 7.03 | 49.70% | 50.30% | 1,926 | 1,129 | 101 | 0 | 615 | 0 | 49 | 32 | 1 |
| Census Tract 7.04 | 53.80% | 46.20% | 4,582 | 2,273 | 425 | 6 | 1,339 | 0 | 423 | 116 | 12 |
| Census Tract 8.02 | 52.90% | 47.10% | 6598 | 4,211 | 353 | 16 | 1,039 | 0 | 793 | 186 | 33 |
| Census Tract 8.03 | 42.00% | 58.00% | 6257 | 4,074 | 553 | 7 | 1,241 | 0 | 113 | 269 | 38 |
| Census Tract 101.03 | 43.20% | 56.80% | 3714 | 2,229 | 144 | 1 | 983 | 0 | 110 | 247 | 57 |
| Census Tract 101.04 | 44.50% | 55.50% | 4891 | 3,435 | 110 | 0 | 850 | 0 | 184 | 312 | 96 |
| Census Tract 102.03 | 52.00% | 48.00% | 5174 | 3,632 | 98 | 0 | 1,160 | 0 | 215 | 69 | 33 |

The other team then downloaded the list of sub-neighborhoods and used latitude and longitude measurements to match census tract numbers to sub-neighborhoods so that they can carry out the analysis of media coverage differences across sub-neighborhoods with different demographics.

- **Coverage of Homicides**

Initially, we were asked to look at the Boston Herald data scraped by last semester's team and evaluate whether it can be used for analyzing racial bias with respect to sub-neighborhoods. But since last year's team findings were completely based on neighborhoods, we couldn't use that directly. So to analyze sub-neighborhoods, we decided to scrape the data from the Boston Herald website. We were asked to use the Wayback machine for scraping data from the website, but even after several attempts and changes we couldn't make the script work. Around this same time, we moved classes completely online and were asked to leave the dorms. Because of this chaos, it was difficult to coordinate and reach out to people for help. Due to the delays and disruptions caused by the transition, after discussions with Spark members we decided to move forward with Boston Globe data as it was already scraped and ready to use.

Data Collection and Preprocessing:

To analyze the coverage of homicides with respect to black and white communities, we used articles from the Boston Globe between 2014 and 2018. Scraping of the data was done by the Fall 2019 team. Along with the news articles, we were also provided a separate list of people who were killed each year, including their name, race, age, and gender.

Data scraped by the last semester's team did not specifically contain only the homicide news. So in order to specifically look at homicide coverage, we first created a subset of articles from the larger pool of data, which specifically covered homicide news. To approach this problem we decided to look for articles containing words with specific words like homicide, murder, crime. To get the better coverage we included other similar words as well in the list, for searching for the homicide sub dataset.

The first thing we did was pre-process the JSON files that contained the raw data for the Boston Globe articles. We converted the JSON files into CSV format for better readability in python code. After this our next step was to find the words similar to words "homicide", "murder", and "crime" in the articles present in the data file. For this we first tokenized the complete dataset and removed the stop words using NLTK library and then applied word2vec model from gensim library. Using the 'most_similar' method of the word2vec model we took the top 5 similar words for each of 3 words defined above for every data file. After that, we took the most relevant similar words to homicide returned by word2vec and added them in the list, which we used to scrape the sub dataset of articles for analyzing homicide coverage. One of the other intentions we had to take semantically similar words other than creating the subset of the dataset specific to homicide was to check if a particular race or gender is getting associated with these 3 defined words. But for the given dataset we didn't find any race or gender in the top 5 associated words with these keywords. Now to create a dataset specific to homicide, we wrote the code to search these semantically similar words across all the articles in each year. While creating a list of keywords for creating the homicide dataset we gave importance to the similarity of word to homicide rather than its larger number of occurrences. Once we got the list of keywords relevant for homicide coverage specific to a particular year, we created a subset of articles for that year containing these keywords. Once this pre-processing was done we had the articles specific to homicide coverage from 2014-2018.

After this, we had to count the number of explicit mentions of a person's name in the homicide dataset. Since all the keywords which were used to create a subset of the homicide dataset were single words, we were able to easily check them without using any specific library to match the keywords. But to check the implicit mention of names we had to match the phrase of 2 words (First Name Last Name). To do this we didn't tokenize the homicide data as it would have created separate tokens for first name and

last name. Instead we used PhraseMatcher of spacy library to match the exact name of the person in the homicide dataset. This code gave us the number of articles a particular person is mentioned explicitly by name. We stored these articles in a separate txt file which mentioned the name of the person explicitly. These txt are created for black and white community separately for every year.

Topic Modeling:

Once we had the files for each race of a particular year, we ran the code to see which are the most relevant frequent words used in the complete set of articles for that race in a particular year. To do this we used the LDA topic modeling method of gensim. As we only wanted the relevant words, we first removed the common stopwords and extended the stopwords list to include words which we found are not relevant to the objective. After every run of LDA we checked the results for the words returned and when we found any word which is generic(e.g. 'rather','tell' etc) we included those in the extended stopwords list. This extended list of stopwords was finalized after many iterations. Once stop words were removed, we also made sure to remove @, newline characters and quotes from the articles and stored articles in a list. To further preprocess the data we formed the bigram and trigrams using the Phrases model of gensim. This is passed to the Phraser function of gensim for efficiency in speed of execution. Next, we lemmatized each word to its root form. We created the NER model through spacy for keeping only nouns, adjectives, verbs and adverbs. We kept only these POS tags because they are the ones contributing the most to the meaning of the sentences. Now to build the LDA topic model using LdaModel() we needed the corpus and the dictionary. We used gensim.corpora Dictionary to create a dictionary for our model and using gensim doc2bow we created a corpus from this dictionary for our model. Once we had a corpus and dictionary we ran the LDA model and tuned various parameters of the model to get most accurate results. After several iterations, we realized that for the given dataset creating 4 topics was giving better results. To visualize the results returned by the LDA model we used two methods: Word Counts of topic keyword and pyLDAvis. Once we ran this code for both the races for every year we realized that as the number of victims belonging to white race are very less we are not able to get relevant words for white community separately as their articles counts were also less. So upon discussion with concerned parties, we merged the data for all 5 years for black and white communities separately and then performed topic modeling on the merged data.

- **Explicit coverage of Black People:**

Data Collection and preprocessing:

We used the same Boston Globe dataset for analyzing the explicit coverage of black people which we used for coverage of homicides. Our initial preprocessing for this part

is the same as that of homicides where we converted data into csv and used the spacy's PhraseMatcher. To check the explicit mention of race in the articles we passed the list of keywords related to race in Phrasematcher.

We used the following key words which were approved by Tanisha Sullivan, the president of the Boston chapter of the NAACP and take into consideration demographic trends that are specific to the Boston area, such as the presence of a sizable Haitian community:

[Black, Afro Latino, Afro Latina, Cape Verdean, Haitian, Haitian-American, African, African-American, Caribbean, Jamaican, Haitian, Dominican, West Indian]

We searched for these keywords on their own, as well as the keywords alongside any of the following words:

[people, male, female, man, men, woman, women, child, kid, youth, community, neighborhood, business, company]

While implementing this part we not only took the count of articles mentioning the race explicitly, we also created the separate dataset of these articles in the corresponding years. We will be using these subsets of articles for sentiment analysis.

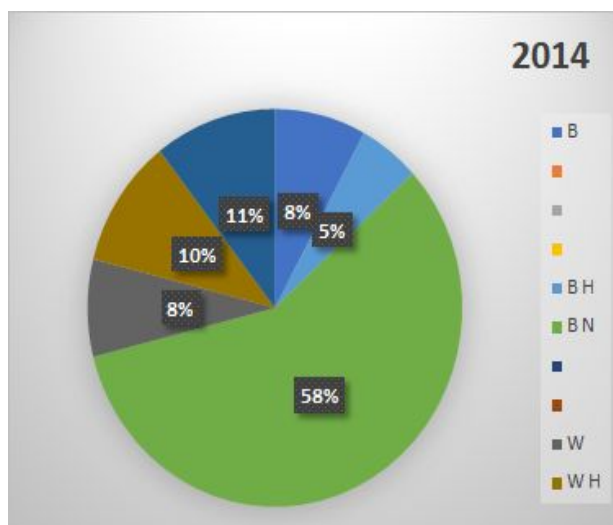
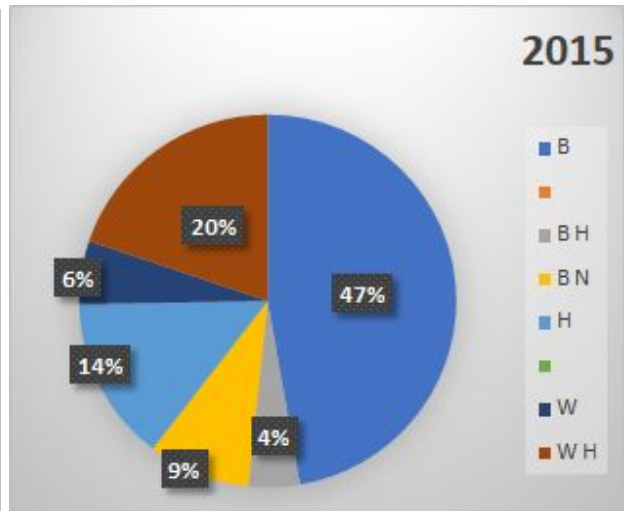
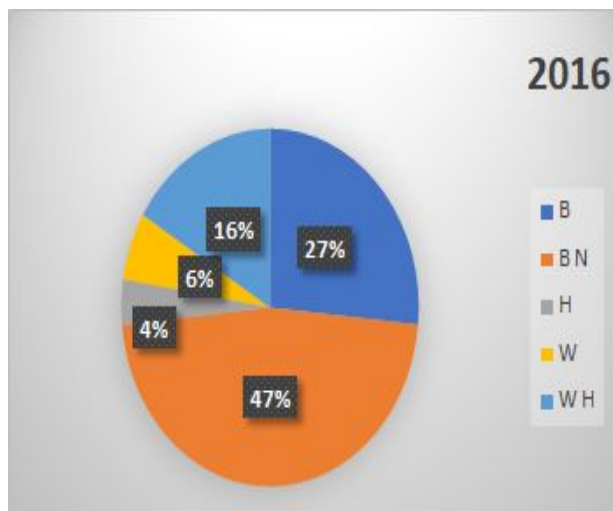
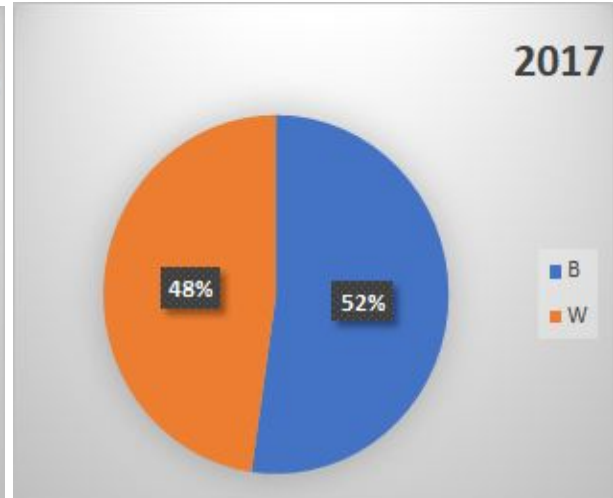
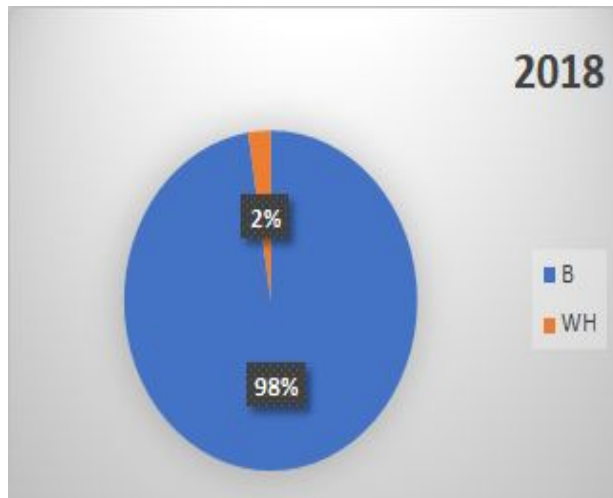
Topic Modeling:

We used the same LDA method implemented for homicide analysis to analyze the common topics across articles that explicitly mention race. We initially tried combining the text files with articles that explicitly mention race for all years together, but unfortunately due to a memory limitation we kept getting an error while running the code. We requested access to the SCC server on campus, but the same error was still showing up so we did the analysis using each year separately.

III. Findings & Observations & Results:

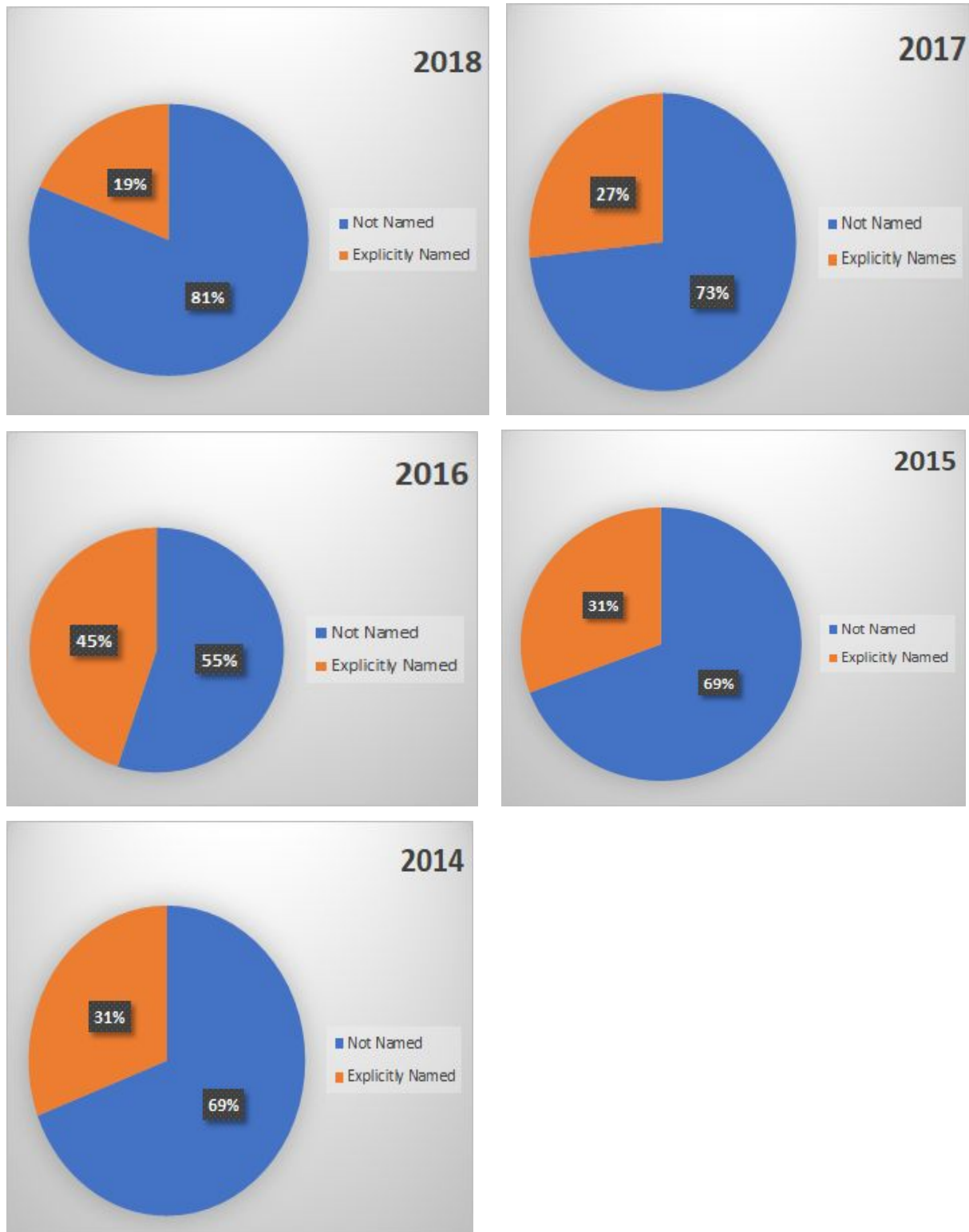
- **Coverage of Homicides**

After comparing the number of articles covering black and white communities for every year separately we observed that most of the names explicitly mentioned are of the black community members. Out of different non-white races, such as Black and Hispanic, Black people were more likely to be explicitly mentioned in the articles.



Results of comparing mentions of explicit names of black and white homicide victims

We also observed that among black people which were provided in the list only less than 50% names were actually explicitly mentioned in the articles.



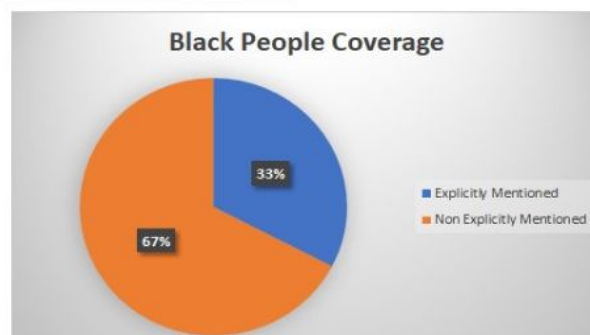
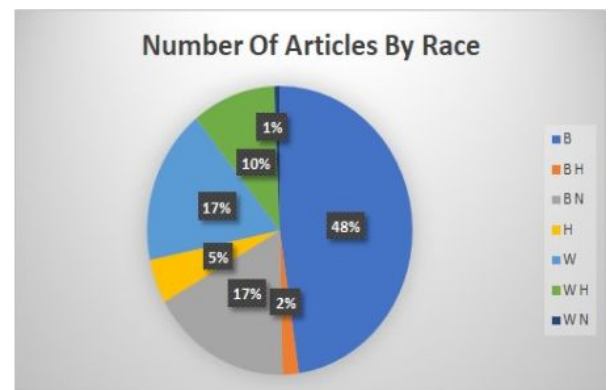
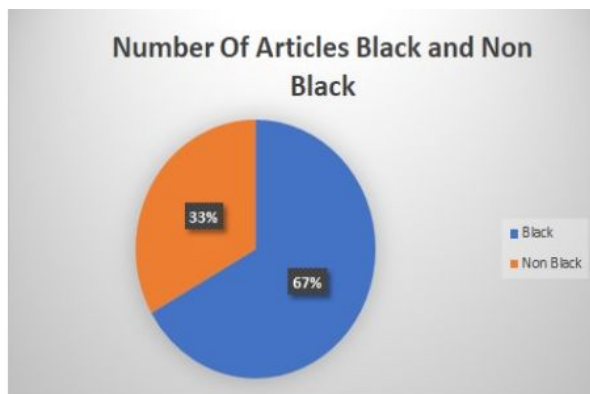
Results for Black people explicitly name vs Black people not named

But as we realized that the number of people murdered from the black community is very high in comparison to white community causing huge imbalance in data to compare the number for every year separately, we decided to combine the data for all 5 years and analyzed the results.

| First Name | Last Name | Year | Race | Gender | Age | Number of Articles | | | |
|---------------|------------|------|------|--------|-----|--------------------|---|----|--|
| Augusta | Mims | 2014 | B | M | 30 | 1 | } | | |
| Christopher | Borgella | 2014 | B | M | 22 | 2 | | | |
| Juan | Nunez | 2014 | B H | M | 21 | 1 | | | |
| Santo Alcadio | Bernabel | 2014 | B H | M | 69 | 1 | | | |
| Shannon | Richardson | 2014 | B N | M | 25 | 1 | | | |
| Shongi | Fernandes | 2014 | B N | M | 24 | 1 | | | |
| Raheem | Robertson | 2014 | B N | M | 21 | 1 | | | |
| Shervonne | Tucker | 2014 | B N | F | 35 | 1 | | | |
| Craig | Marrirro | 2014 | B N | M | 48 | 2 | | | |
| Evens | Archer | 2014 | B N | M | 17 | 2 | | | |
| Emmanuel | Jeudi | 2014 | B N | M | 29 | 3 | | | |
| Dawnn | Jaffier | 2014 | B N | F | 26 | 11 | | | |
| Sherry | Bradley | 2014 | W | F | 32 | 2 | } | | |
| Ryan | Morrissey | 2014 | W | M | 17 | 1 | | | |
| Adrian | Gonzalez | 2014 | W H | M | 17 | 2 | | | |
| Edward | Lawn | 2014 | W N | M | 68 | 1 | | | |
| Sean | Dwyer | 2014 | W N | M | 29 | 1 | | | |
| Janmarcos | Pena | 2014 | WH | M | 9 | 2 | | | |
| | | | | | | | | 27 | |
| | | | | | | | | | |
| | | | | | | | | 9 | |
| | | | | | | | | | |

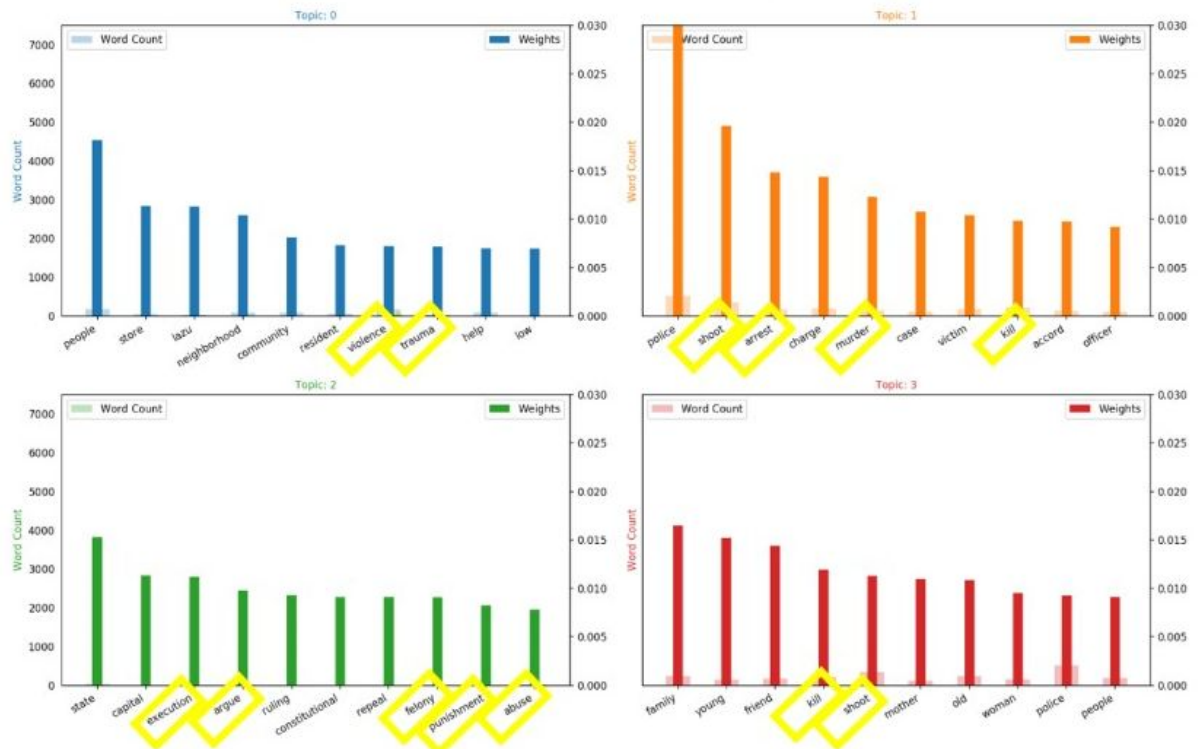
Imbalance in data for black and white community for a single year

Below are the results obtained after merging the all 5 year of data

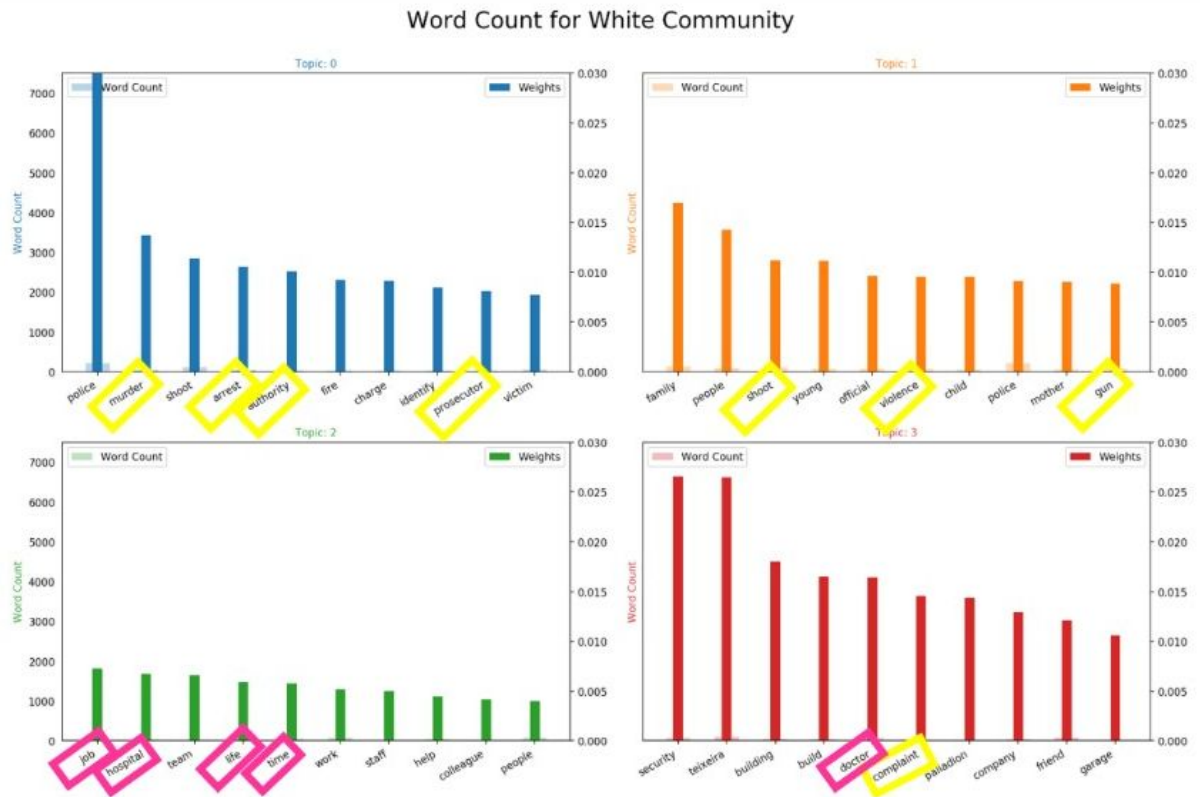


Once we merged the data, we looked for the most relevant frequently used words for black and white community homicide articles.

Word Count for Black Community



Frequently used words for black community



Frequently used words for white community

As we can see from the two charts above, police is the word which has the highest occurrence for both communities. But words which stood out for the black community are : Execution, Felony, Trauma, Abuse. It is also observed that no words with a positive connotation were used frequently in articles mentioning black community homicides, whereas when we observed the words used in the articles for white community we found that there are some positive words like Job, Hospital, Doctor.

On observing both article count and frequently used words, going with articles count will not be a good idea to draw some conclusions due to data imbalance even after combining the data for 5 years. But we can surely say there is some bias against black community on the basis of frequently used words.

- **Explicit Coverage of Black People:**

Below is the result for the volume of a few of the explicit race mentions in 2014 and 2017. The columns indicate the year of coverage. The cells indicate the number of articles in that year that contain the race mentioned in the first cell of that row. For example, in 2017, there were 202 articles that included the phrase “black people”.

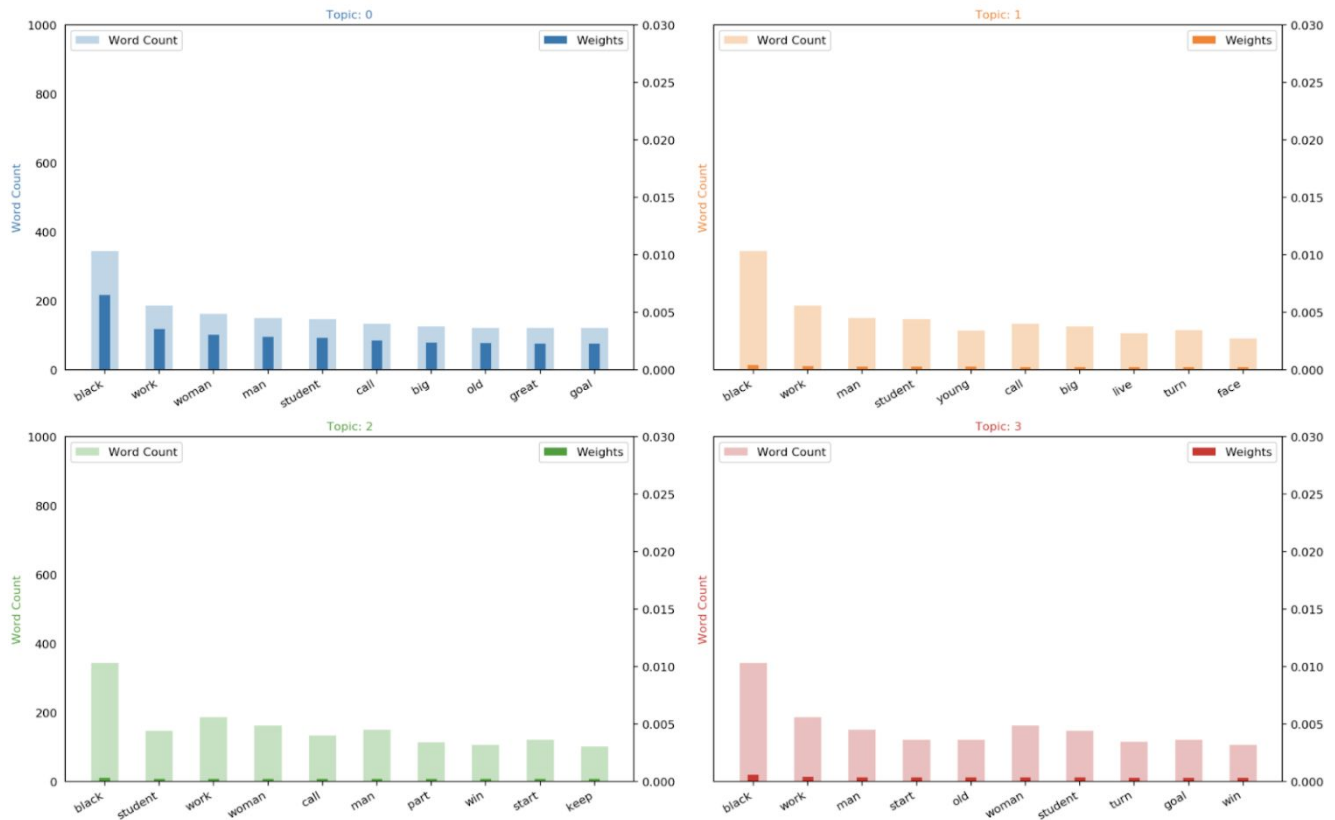
It is worth noting that the volume of race mentions in 2017 is generally larger than in 2014, and this is something that needs to be investigated further. In addition, it is worth

mentioning that our list of keywords for “Black” and for people/individuals is not exhaustive. For instance, there could be mentions of “black doctors” or “black actors” that would count as coverage of Black individuals but might not necessarily be detected by this algorithm since those words are not in the approved list that was provided to us.

| Race Mention | 2014 | 2017 |
|---------------------|-------------|-------------|
| black people | 19 | 202 |
| black male | 4 | 23 |
| black female | 2 | 46 |
| black man | 14 | 223 |
| black men | 15 | 139 |
| black woman | 21 | 101 |
| black women | 7 | 110 |
| black child | 1 | 5 |

For the topic modeling, we repeated several iterations in order to get rid of irrelevant words such as “thing” and “year”. However, even after several iterations, many of the results were still not significant. Part of the reason could be due to the fact that we could only analyze the articles for one year at a time rather than all 5 years together due to the memory allocation restrictions mentioned earlier. For example, below is the graph for the frequently mentioned words for 2016 even after several iterations of eliminating irrelevant words and extending the list of stopwords to be removed. One of the most frequently mentioned words is “black”, which is expected since these are the articles that explicitly mention race. However, none of the other words give any insight regarding any topics or biases present in these articles. Words such as “start” and “keep” are not significant. Therefore, our analysis of explicit mention of race is generally inconclusive due to memory issues when combined all the year data.

Frequently Used Words for Race Mentions



IV. Conclusion

Through this project we got the opportunity to investigate the media coverage of the Boston Globe with regards to different racial groups. We tried to help our client by providing a better understanding of the difference in coverage of homicides in the case of the black community and white community along with providing analysis on explicit mention of race. For homicides, we analyzed volume of coverage as well as the percentage of named and unnamed mentions in articles, and for explicit mentions of race we tried to analyze the topics most commonly present in those articles

This information could be very useful for the NAACP in its mission to fight for social justice and achieve racial equity because it sheds light on some of the disparities experienced by the Black community in terms of media coverage, specifically regarding coverage of homicides.

We want to give special thanks to Ziba and Gowtham for their patience and help.