NAACP Media Research

CS506 - Spark! Project

Ziyi Jiang, lynnj@bu.edu | Naiwen Xu, xnw1127@bu.edu | Yuhao He, hey1@bu.edu | Manyuan Lu,manyuan@bu.edu

I. Summary & Problem Statement

NAACP, the National Association for the Advancement of Colored People, was founded on February 12, 1909. It is the nation's foremost, largest, and most widely recognized civil rights organization. The vision of NAACP is to ensure a society in which all individuals have equal rights without discrimination based on race. Our client, NAACP Boston seeks to understand the coverage of Boston Media in covering Black people and Black communities in Boston.

To accomplish this goal, we are going to define Black communities using census tract level data, collect data from mainly Boston Globe, analyze the content and sentiment using Natural Language Processing tools. Our analysis will compare the difference between Black people related articles and non-Black people related ones. In addition, we will identify the frequent keywords in the news about Black people from the Boston Globe. In doing so, we can hopefully give our client a better picture of the coverage of news about Black community and thus eliminate the race discrimination.

II. Methodology & Algorithms

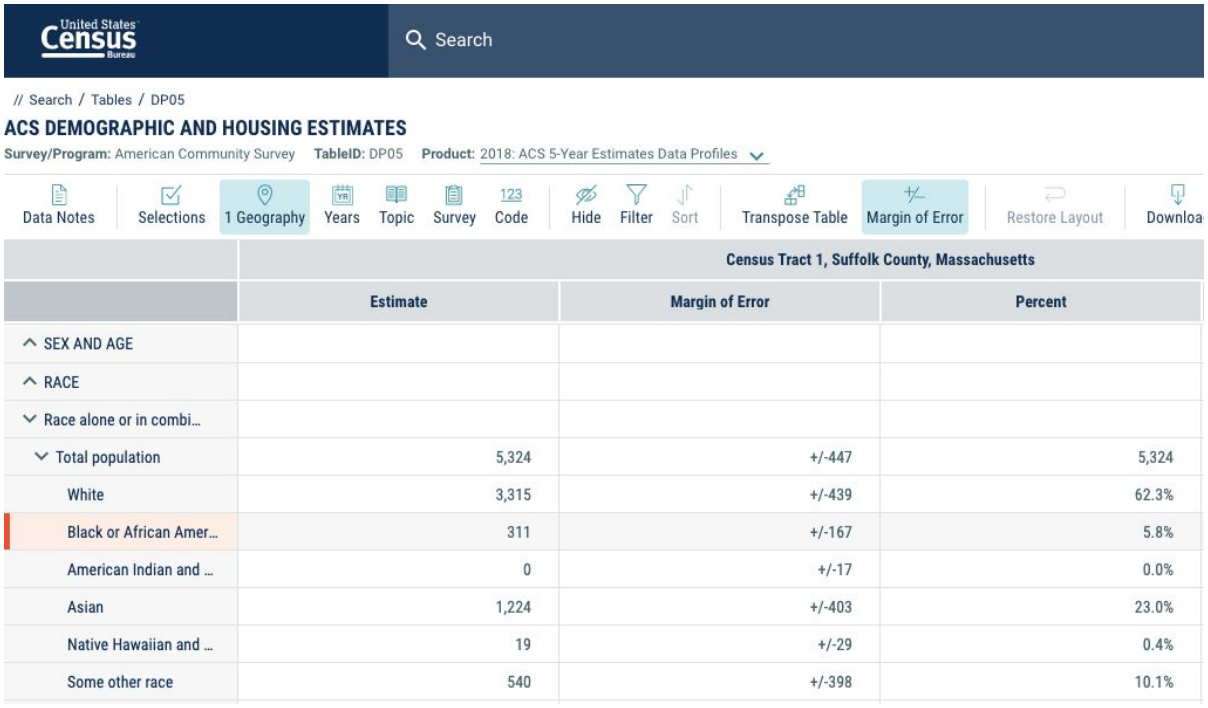
Step 1: Collecting data

As we plan to focus on analyzing the news from the angle of how Black communities are covered based on sub-neighborhoods, the first step is to define which neighborhoods can be considered as Black or non-Black communities. To do so we collected the list of sub-neighborhoods from the United States Geological Survey (see Table 1) and the demographics for each census tract from the United States Census Bureau (see Table 2). Then we used a look-up tool, Census Batch Geocoder which will return the corresponding census tract code if we enter a specific address within one neighborhood, to determine the sub-neighborhoods with the largest concentration of Black people.

Feature Name	ID	Class	County	State	Latitude	Longitude	Ele(ft)	Map	BGN Date	Entry Date
Charlestown	619308	Populated Place	Suffolk	MA	422240N	0710343W	43	Boston North	-	24-FEB-1974
Chelsea	612723	Populated Place	Suffolk	MA	422330N	0710158W	39	Boston North	-	24-FEB-1974
East Boston	612734	Populated Place	Suffolk	MA	422230N	0710221W	13	Boston North	-	24-FEB-1974
Franklin Park	612747	Populated Place	Suffolk	MA	422630N	0710058W	36	Boston North	-	24-FEB-1974
Orient Heights	612795	Populated Place	Suffolk	MA	422315N	0710013W	16	Boston North	-	24-FEB-1974
Revere	612810	Populated Place	Suffolk	MA	422430N	0710043W	10	Boston North	-	24-FEB-1974
Ashmont	612861	Populated Place	Suffolk	MA	421700N	0710408W	82	Boston South	-	24-FEB-1974
Back Bay	617071	Populated Place	Suffolk	MA	422100N	0710513W	13	Boston South	-	24-FEB-1974
Bay Village	1970982	Populated Place	Suffolk	MA	422102N	0710411W	16	Boston South	-	15-JAN-2003
Beacon Hill	619512	Populated Place	Suffolk	MA	422131N	0710404W	92	Boston South	-	27-AUG-2002
Boston	617565	Populated Place	Suffolk	MA	422130N	0710335W	46	Boston South	07-OCT-1931	24-FEB-1974
Cedar Grove	612891	Populated Place	Suffolk	MA	421650N	0710328W	23	Boston South	-	24-FEB-1974
Chinatown	606756	Populated Place	Suffolk	MA	Unknown	Unknown	-	Boston South	-	01-JAN-1992
City Point	612897	Populated Place	Suffolk	MA	422010N	0710143W	26	Boston South	-	24-FEB-1974
Dock Square	1877489	Populated Place	Suffolk	MA	422107N	0710326W	13	Boston South	-	17-AUG-2000

row(s) 1 - 15 of 73

Table 1. List of sub-neighborhoods from the United States Geological Survey



	Census Tract 1, Suffolk County, Massachusetts		
	Estimate	Margin of Error	Percent
SEX AND AGE			
RACE			
Race alone or in combi...			
Total population	5,324	+/-447	5,324
White	3,315	+/-439	62.3%
Black or African Amer...	311	+/-167	5.8%
American Indian and ...	0	+/-17	0.0%
Asian	1,224	+/-403	23.0%
Native Hawaiian and ...	19	+/-29	0.4%
Some other race	540	+/-398	10.1%

Table 2. Demographics for each census tract from the United States Census Bureau

After defining the sub-neighborhoods, the next step is to collect the articles between Year 2014 and Year 2018 from major Boston news outlets. Our team tried to scrape the articles from Boston Globe using the Wayback machine API for Scrapy, but we encountered some problems like anti-scraping techniques that will keep their websites being scraped. Thus, by getting permission, we decided to use the existing data collected by a team who took this project during the last semester.

Step 2: Filtering News

We initially planned to filter news by the addresses mentioned in the article. However, it turns out that it might be too specific and we can not ensure if the whole article is really about that place, or just a location it happened to mention. Then, we moved to a new mechanism – just look up the keywords of the sub-neighborhood it mentioned to determine whether it is about a black community. If an article doesn't mention any neighborhood, we simply drop that piece of article. If more than three neighborhoods

except Boston are mentioned in the article, we also drop it as we can consider the article as not focusing on talking about any neighborhood. Moreover, if the article only contains a single location ‘Boston’, we also remove it from our data because it may point to the Greater Boston Area instead of the neighborhood Boston.

To determine whether a sub-neighborhood is a black community, we calculated the black proportion according to the demographics we collected from the United States Census Bureau. If the black proportion of a sub-neighborhood is greater than 50%, we consider it as a Black neighborhood; otherwise, we labeled it as a non-Black neighborhood.

	Sentence	Score	Result	Neighborhood	Sub-neighborhood	Black or Not [Y/N]
0	'', ' individuals who worked in finance, insu...	{'neg': 0.029, 'neu': 0.804, 'pos': 0.167, 'co...	Positive	[Roxbury]	[Boston, Roxbury]	N
1	'Dr. Kenneth C. Edelin, whose historic 1975 m...	{'neg': 0.077, 'neu': 0.837, 'pos': 0.085, 'co...	Positive	[Roxbury]	[Boston, Roxbury]	N
2	'When they go up, they form the centerpiece o...	{'neg': 0.063, 'neu': 0.883, 'pos': 0.054, 'co...	Negative	[Roxbury]	[Boston, Roxbury]	N
3	'First Night Boston almost didn't happen this...	{'neg': 0.019, 'neu': 0.864, 'pos': 0.118, 'co...	Positive	[Back Bay]	[Back Bay, Boston]	N
4	'A woman crossing the Meridian Street Bridge ...	{'neg': 0.087, 'neu': 0.843, 'pos': 0.07, 'com...	Negative	[East Boston, Chelsea]	[Chelsea, East Boston, Boston]	N
...	...	...	...	...	...	...
3198	'Are you hoping to enjoy some Peking duck or ...	{'neg': 0.004, 'neu': 0.911, 'pos': 0.086, 'co...	Positive	[Allston]	[Boston, Allston, Chinatown]	N
3199	'Dozens of taxi drivers and industry supporte...	{'neg': 0.056, 'neu': 0.883, 'pos': 0.061, 'co...	Negative	[Dorchester]	[Boston, Dorchester]	N
3200	'In fireworks and ice sculptures, through pup...	{'neg': 0.084, 'neu': 0.809, 'pos': 0.108, 'co...	Positive	[Back Bay, East Boston, South Boston, South End]	[East Boston, Back Bay, Boston, South Boston]	N
3201	'', 'Chrystal Kornegay, a Democrat who leads ...	{'neg': 0.026, 'neu': 0.91, 'pos': 0.064, 'com...	Positive	[Jamaica Plain, Roxbury, Chelsea]	[Chelsea, Jamaica Plain, Roxbury]	N
3202	'', 'Governor-elect Charlie Baker has tapped ...	{'neg': 0.025, 'neu': 0.88, 'pos': 0.095, 'com...	Positive	[Beacon Hill, Dorchester, South Boston]	[Beacon Hill, Boston, Dorchester, South Boston]	N

Table 3. Categorizing the news as Black community related or not

Step 3: Sentiment Analysis

We used a python package called *VADER-Sentiment-Analysis* as a tool for sentiment analysis. It computed a compound score by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive), which is a normalized, weighted composite score. We can use this score to determine if an article is overall

positive or negative.

	Articles	Score	Result
0	, 'WASHINGTON (AP) ü United and eager to respond to a	{'neg': 0.074, 'neu': 0.811, 'pos': 0.115, 'compound': 0.9928}	Positive
1	'The five Democratic candidates for governor of Massach	{'neg': 0.057, 'neu': 0.882, 'pos': 0.061, 'compound': -0.6517}	Negative
2	'Given that multiple television outlets were interested in h	{'neg': 0.023, 'neu': 0.916, 'pos': 0.061, 'compound': 0.9299}	Positive
3	'PORTLAND, Maine ü In 1910, Georges Braque and Pablo	{'neg': 0.045, 'neu': 0.874, 'pos': 0.08, 'compound': 0.9746}	Positive
4	'A coalition of Massachusetts health care providers that c	{'neg': 0.019, 'neu': 0.86, 'pos': 0.121, 'compound': 0.999}	Positive
5	'\n', 'Saying itüs worried that mounting medical costs are	{'neg': 0.026, 'neu': 0.862, 'pos': 0.111, 'compound': 0.9985}	Positive
6	'\n', ' parents, high school graduation ceremonies bring a	{'neg': 0.068, 'neu': 0.801, 'pos': 0.13, 'compound': 0.9866}	Positive
7	'\n', 'publication Tuesday of Hillary Clintonüs üHard Choic	{'neg': 0.148, 'neu': 0.784, 'pos': 0.068, 'compound': -0.996}	Negative
8	, 'WASHINGTON ü In one of the most stunning primary e	{'neg': 0.075, 'neu': 0.833, 'pos': 0.092, 'compound': 0.9218}	Positive

Table 4. Vader-Sentiment-Analysis results

We run *Vader-Sentiment-Analysis* on both black and non-black related articles, so we can analyze the difference between the sentiments of these two sets of articles.

Step 4: Getting Significant Topics

The strategy we used to identify the significant topics is *Latent Semantic Analysis.* It is an efficient way of analyzing the text and finding the hidden topics by understanding the context of the text. We run *Latent Semantic Analysis* on all news about black neighborhoods during each year so that we can have an idea of the main topics about black neighborhoods reported in the Boston Globe.

III. Findings & Observations

Data:

We've collected reliable demographics data to determine the Black and non-Black neighborhoods. We also collected all news between 2014 and 2018 from the Boston Globe, in which there are 192,400 pieces of news collected in total and they are classified by year from 2014 to 2018.

Black Proportion:

The distribution of Black Population is highly right skewed. Among 205 census tracts in Suffolk County, about 90 of them have less than or equal to 10% of African American population. Few of the census tract have more than 50% African American residents. According to this fact, we can see that Suffolk County does not have a lot of “Black American Community”; the Great Boston Area is not densely populated with black American residents.

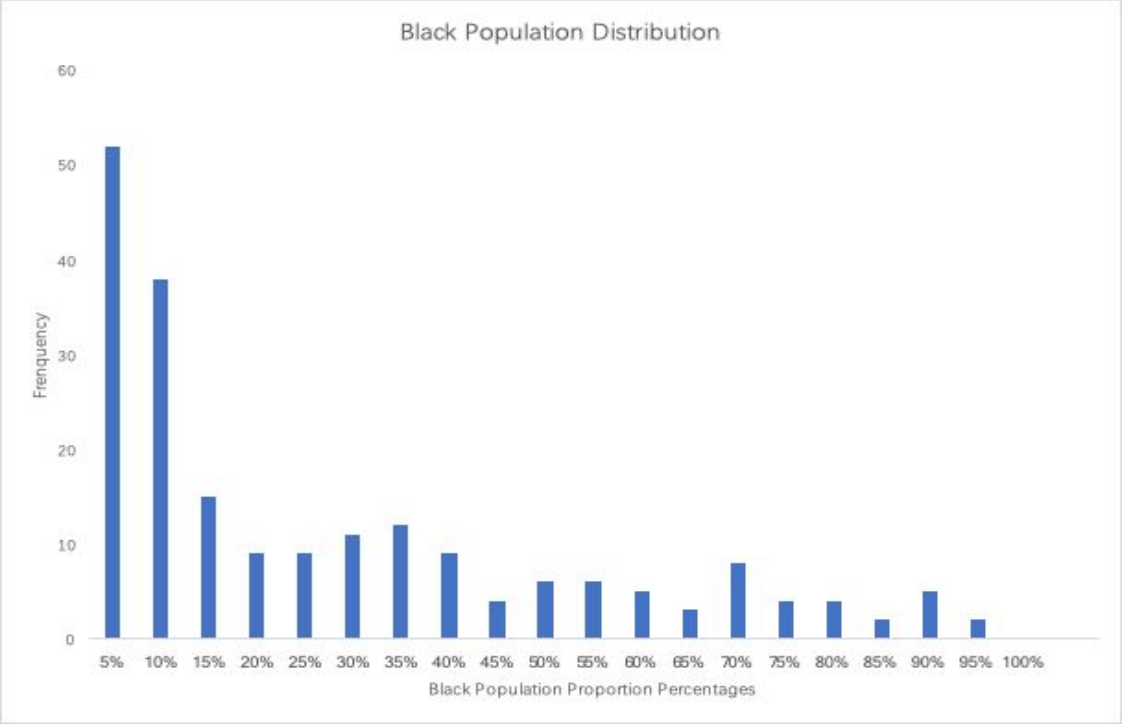


Fig 1. Black Proportion Distribution

Coverage:

We used only neighborhood names to calculate the coverage. For Boston Globe, the coverage (number of articles about black neighborhoods/number of articles about all neighborhoods) is 9.5% and hasn’t changed much during the past five years (see Table 3). As we can see, using only neighborhood names to determine the coverage of black people is not enough.

Dataset	# News about black nbh	# News about all nbh	# News	Coverage
globe2014	307	3203	18576	0.09584764
globe2015	416	4474	27000	0.09298167
globe2016	514	5476	35562	0.09386413
globe2017	643	6718	41876	0.09571301
globe2018	1112	11495	69386	0.09673771

Table 4. Statistics of Coverage

Sentiment Analysis:

The pie charts below show the sentiment of articles for white and black neighborhoods from 2014 to 2018. The red area represents the proportion of positive news about white neighborhoods in each year. The purple area represents the proportion of negative news about white neighborhoods in each year. The blue area represents the proportion of positive news about black neighborhoods. The orange area represents the proportion of negative news about black neighborhoods.

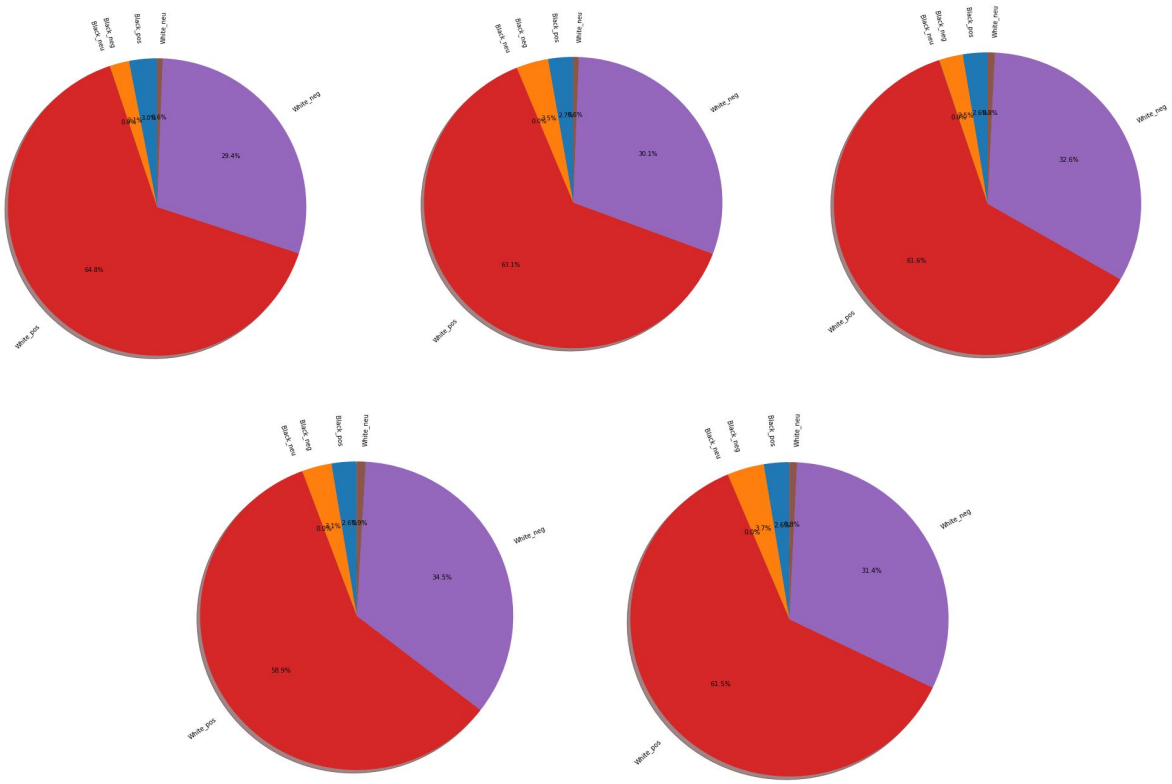


Fig 2. The sentiment of articles for white and black neighborhoods from 2014 to 2018

Since the percentage in the pie charts cannot be seen clearly, we summarized the statistics and made a table below.

Column1	2014	2015	2016	2017	2018
white_positive	65.40%	63.10%	61.60%	58.90%	61.50%
white_neutral	0.50%	0.60%	0.80%	0.90%	0.80%
white_negative	29.00%	30.10%	32.60%	34.50%	31.40%
black_positive	2.00%	2.70%	2.60%	2.60%	2.60%
black_neutral	0%	0%	0%	0%	0%
black_negative	3.10%	3.50%	2.50%	3.10%	3.70%

As we can see from the pie charts and table above, the overall pattern for the attitude toward Black and White neighborhood is roughly the same. However, the percentage of

negative attitudes toward black neighborhoods is increasing from 2015 to 2018. It dropped a little bit in 2016, but the percentage of negative reports toward black neighborhoods surged back and even higher than it was in 2014.

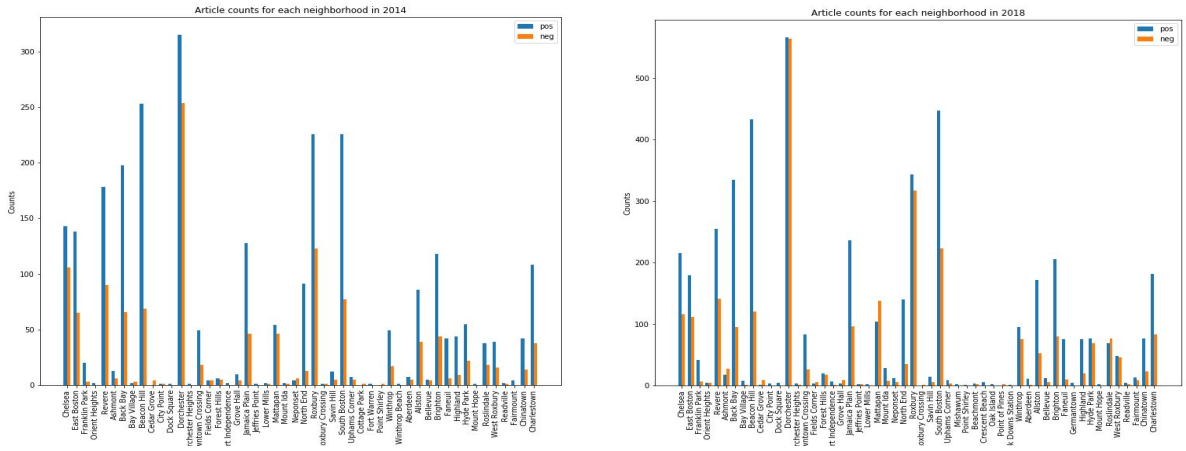


Fig 3. The sentiment of articles for sub-neighborhoods in 2014 and 2018

The bar charts above are the sentiment of articles for sub-neighborhoods. We remove Boston from the chart, because when people talk about Boston, it might refer to a different scale of area in different contents, and it has way more articles than others. Therefore, we considered it as an outlier and removed it from the images. From the charts, we can see that people pay a lot of attention to Dorchester, which is a black neighborhood. However, compared to other sub-neighborhood like Beacon Hill or Back Bay, Dorchester seems to gain more negative reports than those white neighborhoods. Other typical black sub-neighborhood, such as Ashmont, Roxbury and Mattapan, their negative reports are all equal to or greater than their positive reports.

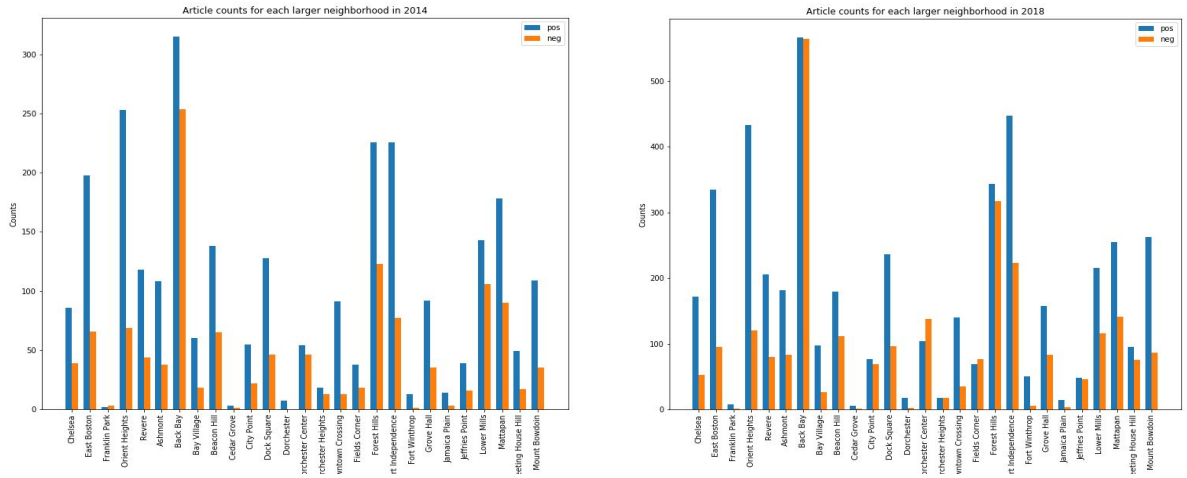


Fig 4. The sentiment of articles for larger neighborhoods in 2014 and 2018

For larger neighborhoods, the sentiment of articles had less connection with whether the area is predominantly balck or white. It became more righteous, and there is no tendency for negative reports as we can find in sub-neighborhoods.

Topics:

2014	[‘school’, ‘police’, ‘community’, ‘people’, office’]
	[‘school’, ‘student’, ‘police, ‘high’, ‘madison’, ‘office’, ‘park’, ‘education’, ‘vocation’]
	[‘police’, ‘school’, ‘church’, ‘office’, ‘work’, ‘people’, ‘job’, ‘student’]
	[‘station’, ‘radio’, ‘police’, ‘community’, ‘gun’]
	[‘church’, ‘police’, ‘bank’, ‘groover’, ‘job’, ‘loan’, college’, ‘street’]
2015	[‘school’, ‘police’, ‘student’, ‘college’, ‘office’, ‘high’, ‘shoot’]
	[‘school’, ‘student’, ‘police’, ‘college’, ‘build’, ‘high’, ‘square’, ‘resident’, ‘develop’]
	[‘bond’, ‘police’, ‘office’, ‘bella, ‘mccarthy’, ‘taylor’, ‘depart’, ‘harrison’]
2016	[‘school’, ‘police’, ‘student’, ‘start’, ‘office’, ‘high’, ‘later’, ‘time’, ‘charter’]
	[‘police’, ‘office’, ‘school’, ‘work’, ‘shoot’, ‘companies’, ‘job’]
	[‘water’, ‘charter’, ‘test’, ‘innovation’, ‘lead’, ‘fountain’, ‘level’, education’]
2017	[‘school’, ‘student’, ‘metco’, ‘police’, ‘district’, ‘teacher’, ‘state’, ‘education’]
	[‘jackson’, ‘police’, ‘winner’, ‘walsh’, ‘mayor’, ‘properties’, ‘build’, ‘girl’]

	[‘fire’, ‘tito’, ‘depart’, ‘owner’, ‘black’, ‘neighbor’]
‘2018	[‘school’, ‘student’, ‘police’, ‘college’, ‘high’, ‘office’ ‘boston’, ‘class’, ‘shoot’]
	[‘school’, ‘police’, , ‘office’, ‘student’, ‘boston’, ‘people’, ‘build’, ‘work’]
	[‘police’, ’office’, ‘bond’, ‘boston’, ‘black’, ‘bella’, depart’, ’train’]
	[‘train’, ‘bond’, ‘line’, ‘mbta’, ‘bella’, ‘station, ‘street’]

The table above contains a list of the top unique words discovered in each year by our *Latent Semantic Analysis*. Based on the information above, we can summarize that the significant topics reported about black neighborhoods do not change much from 2014 to 2018 and they are mainly about education, criminals, employment, and livelihood.

IV. Conclusions

This project aims to help our client NAACP, the National Association for the Advancement of Colored People, to have a better idea of the coverage of Boston Media in covering Black people and Black communities in Boston. We categorized articles into Black community related and non-black related by identifying neighborhood keywords in articles, and we analyzed sentiments associated with each category. As a result, we could assess how the coverage and the sentiment of news has changed from 2014 to 2018. We also figured out the most common topics of stories about Black neighborhoods and sub-neighborhoods. These findings can hopefully help our clients explore Boston Media’s current situations.

The findings show that people pay more attention to the black neighborhoods as the coverage of black news increased a little from 2014 to 2018. However, the number of negative news about black neighborhoods increased from 2014 to 2018, and the black neighborhoods generated more negative reports proportionally to the white neighborhoods. One possible reason for such a phenomenon is the stereotype of black neighborhoods. In addition, after identifying the significant topics, it is worth noting that the topics about black neighborhoods have some words related to criminals and

employment. This also reflects that the sentiment of news about black people tends to be negative.

In the process of research, we found that the most difficult part is to link each census tract to its corresponding sub-neighborhood. Since there is no official record for the sub-neighborhood geographic area, our team had to collect indirect data from various websites to generate the final sub-neighborhood demographic breakdown.
The data can be found and downloaded here:
https://docs.google.com/spreadsheets/d/1eVrKc53fIOwa9NCEz-iJlTJd1-Ro7iO0GrPDPqK1ifo/edit?usp=sharing

The code can be downloaded from here:
https://github.com/lynnjiangsmart/CS506-Spring2020-Projects/tree/NAACP/NAACP-1