Media Analysis of Black Americans and Communities in Boston

NAACP - Boston Chapter

- Rayyan Nasr, Mahmoud Khalil, Stephanie Forbes, Mani Singh, Dingjie Chen, Manish Patel, and Zheng Hui

- December 7, 2020

Outline:

1. Motivation and Background
2. Census Coverage
3. Topic Modeling
4. Crime Coverage
5. Exploring Different Modeling Techniques
6. Sentiment Analysis
7. Entity Recognition
8. Limitations / Future Work

Motivation and Background:

- Evaluate the media coverage of Black Americans in the Boston area over the past five years (2014-2018)
- Coverage includes: Overall coverage, predominantly black neighborhoods and sub-neighborhoods, homicide coverage and more

Previous Challenges:

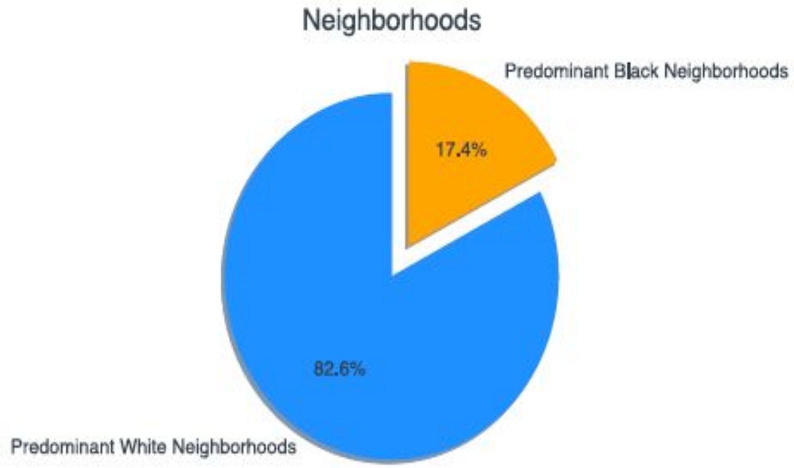
- Not enough specificity with the topics being modeled
- Visualization of topic modeling not clear enough for people with no prior CS background

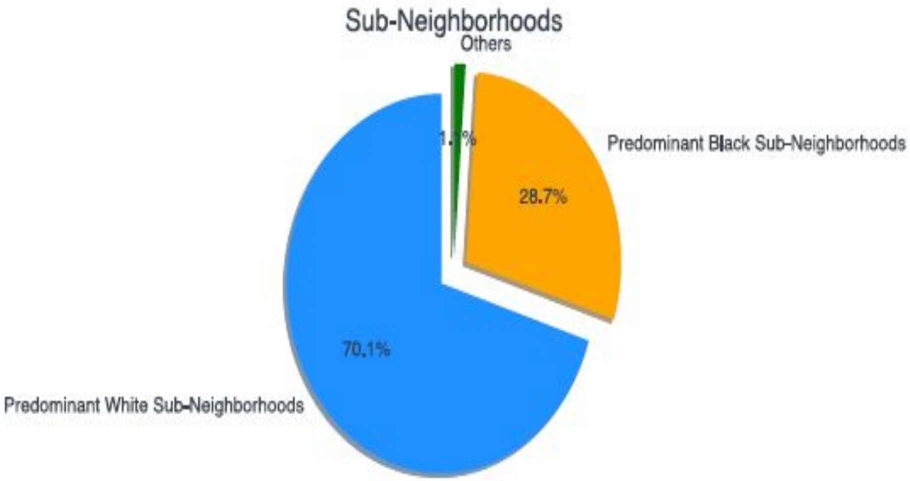
Data Sources:

- Revised Census Data
- WGBH and WBUR articles
  - ❑ Used beautifulsoup to web-scrape years 2014 through 2018
- Homicide Victims in Boston
  - ❑ Also previously collected

Census Data:

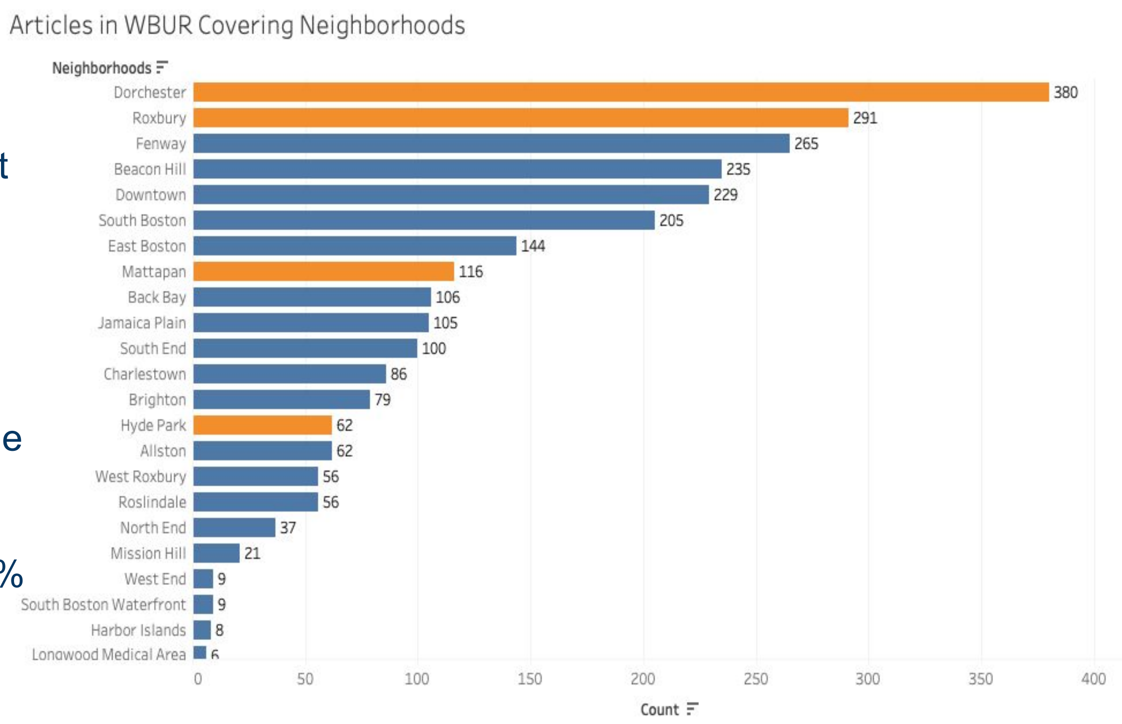
- White Americans constitute 58.5%, while Black Americans 29.5% of the total population
- Remaining 12% constitute of American-Indians and Asians





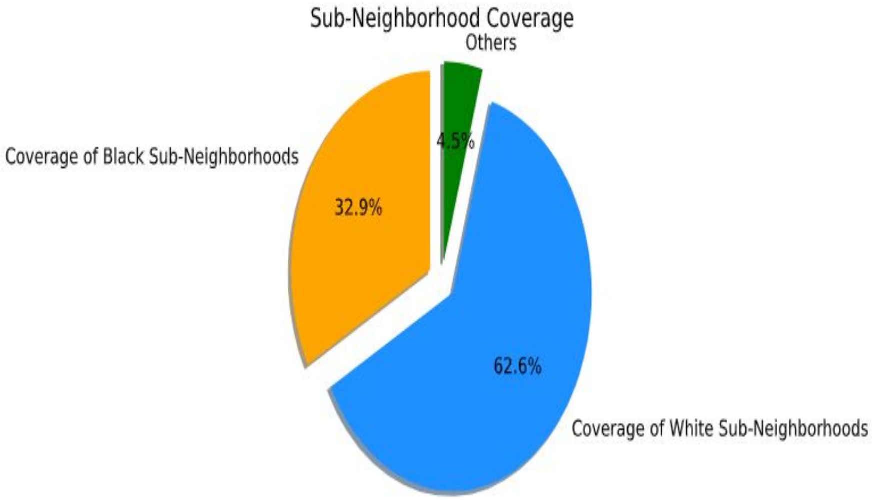
Coverage of Neighborhoods in WBUR:

- Between 2014-2018 there were 12,656 articles
- 1,818 articles covered white predominant neighborhoods
- 849 articles covered black predominant neighborhoods
- The Black population is only 29.5% of the Boston area, but predominantly Black neighborhoods were the subject of 31.8% of news articles that had a geographic mention



Coverage of Sub-Neighborhoods in WBUR:

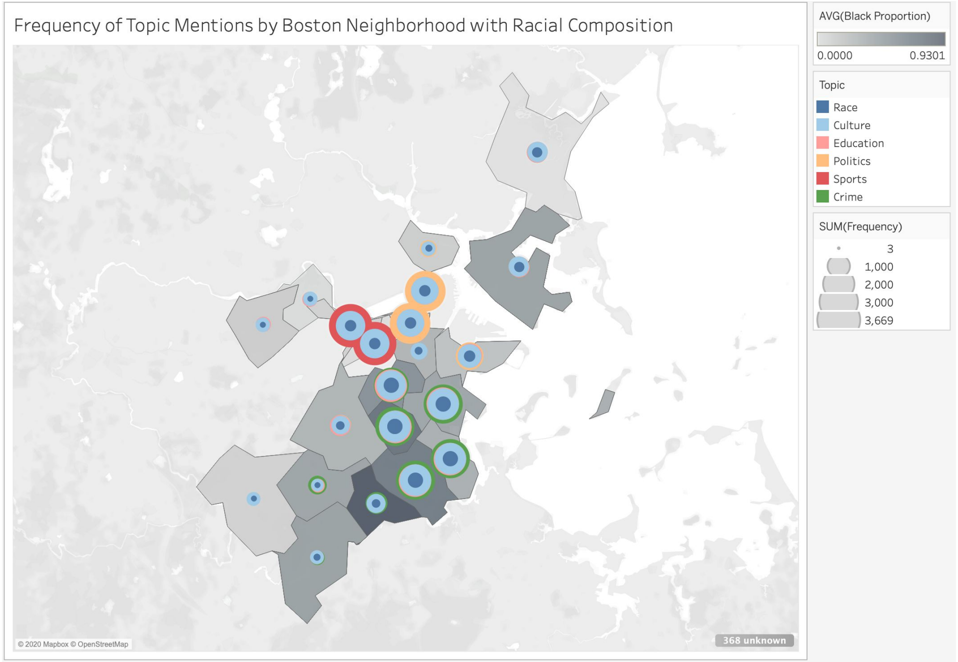
- Out of the 61 predominant white sub-neighborhoods, 33 were covered
- Out of the 25 predominant black sub-neighborhoods, 9 were covered
- 1,673 articles covered white predominant sub-neighborhoods
- 878 articles covered black predominant sub-neighborhoods



Visualization Methods

Using **Tableau** primarily to handle visualizations after preprocessing the data sets:

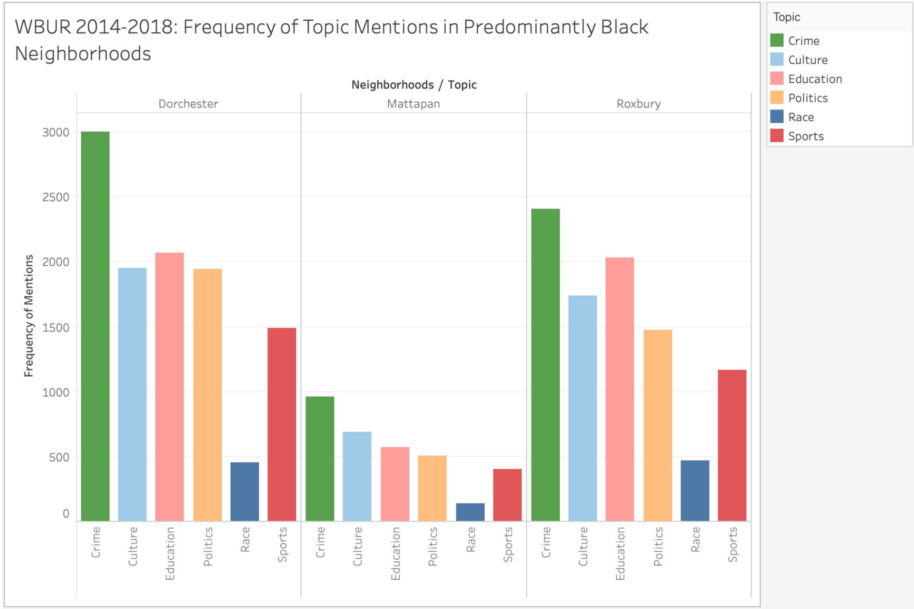
- The racial breakdowns (namely the proportion of the population that is Black) for each neighborhood
- Using the content of articles to establish topics (from a pre-set list) mentioned within each article
- Frequency of topic mentions corresponds to size of point, while color shows the topic



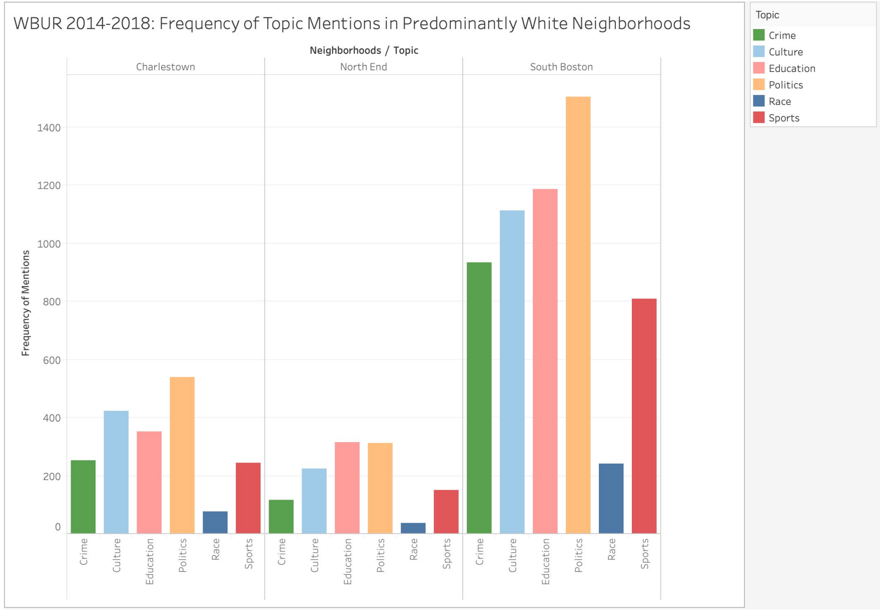
Visualization of Topic Mentions in WBUR Articles from 2014-2018 for Different Neighborhoods of Boston alongside their Racial Distribution

Topic Mentions in Predominantly Black vs. White Neighborhoods

Predominantly Black Neighborhoods



Predominantly White Neighborhoods



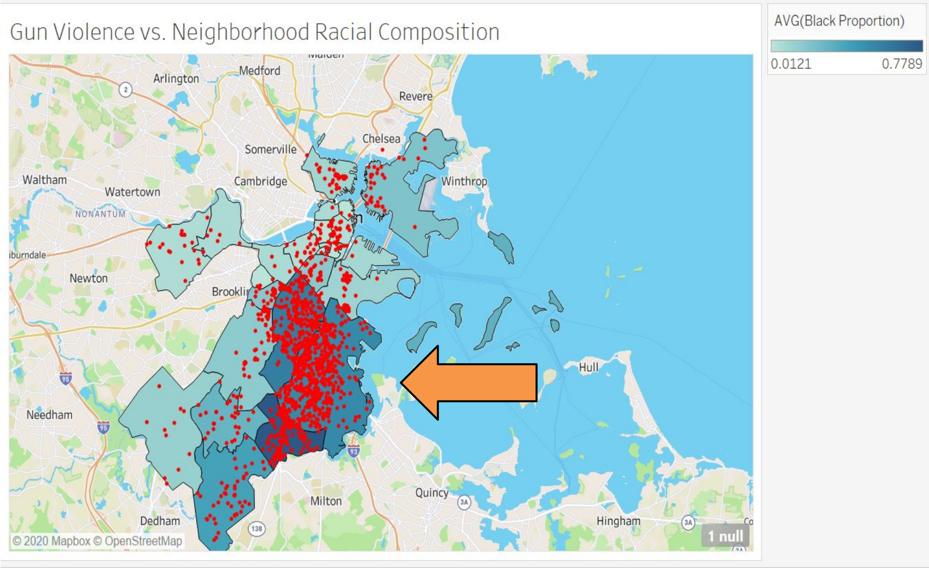
Coverage of Gun Violence: Process

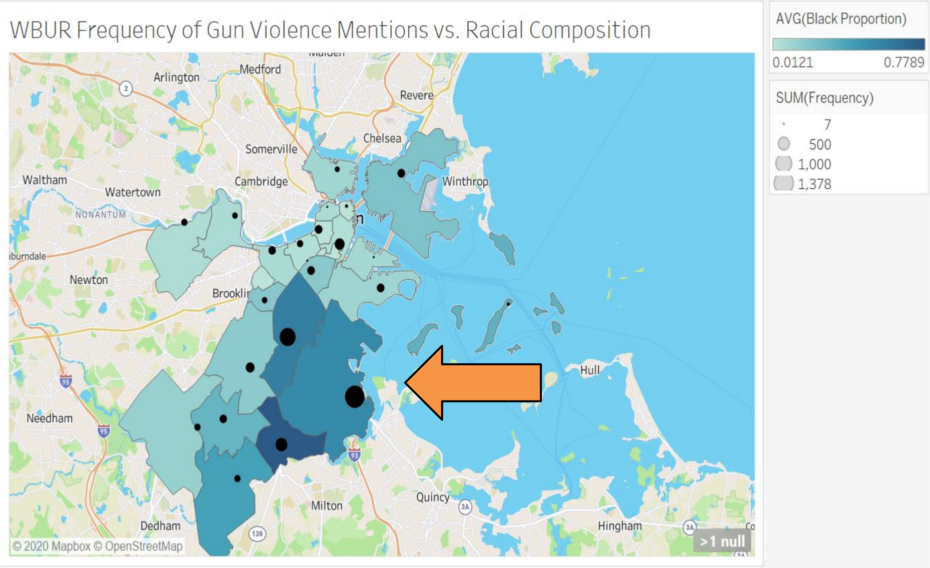
Using a CSV of gun violence incidents (2014-2018), we can identify in which neighborhoods gun violence is most prevalent and if its coverage is proportionally evident:

- Looking for the frequency of gun violence mentions in the articles for each neighborhood
- Geocoding the incidents and visualizing them with article mentions by neighborhood
- If there is a discrepancy, does the racial composition play a role? → bias?

```
import pandas as pd
import geopy
from geopy.extra.rate_limiter import RateLimiter


gv_df = pd.read_csv("gun_violence_incidents_2014-2018.csv")

geolocator = geopy.Nominatim(user_agent="myGeocoder")
geocode = RateLimiter(geolocator.geocode, min_delay_seconds = 1)

gv_df['Full Address'] = gv_df['Address'] + "," + gv_df['City Or County'] + "," + gv_df['State']

gv_df['Location'] = gv_df['Full Address'].apply(geocode)

gv_df['Point'] = gv_df['Location'].apply(lambda loc: tuple(loc.point) if loc else None)
```

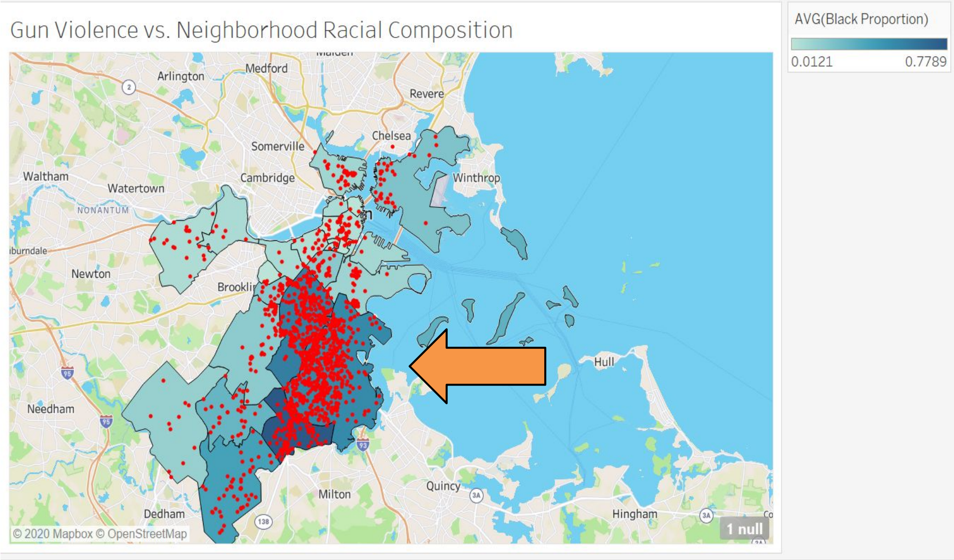
Coverage of Gun Violence: WBUR

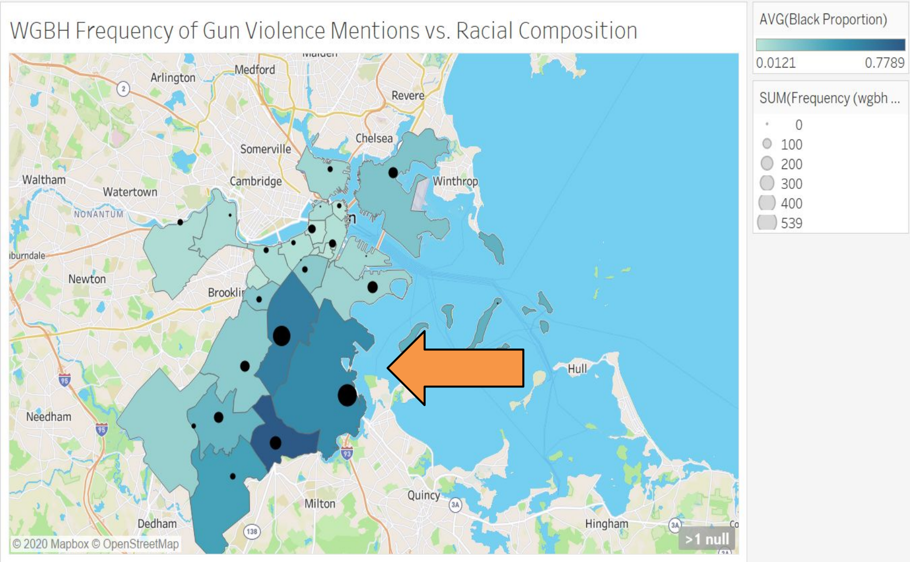




No clear discrepancies here, the dense areas of gun violence incidents have the highest gun violence coverage (except for Matapan possibly)

Coverage of Gun Violence: WGBH





Once again, no clear discrepancies here, as the dense areas of gun violence incidents have the highest gun violence coverage (again except for Matapan possibly)

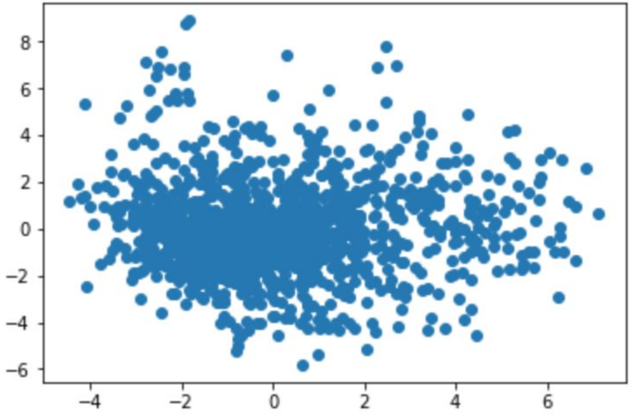
Exploring Different Modeling Techniques

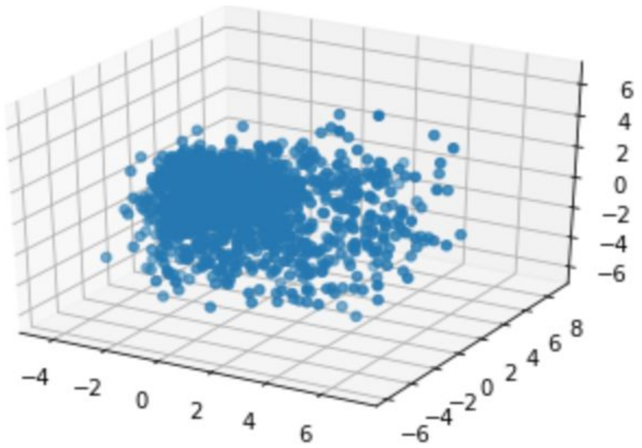
- Previous effort was to perform topic modeling using LDA.
    - LDA is a Bag of Words (BoW) technique.
    - Lose word order.

- Idea: use embeddings to model corpus of documents
    - Technique: Doc2Vec (extension of Word2Vec but for whole document)
    - Capture the ideas embedded within each document.

Exploring Different Modeling Techniques

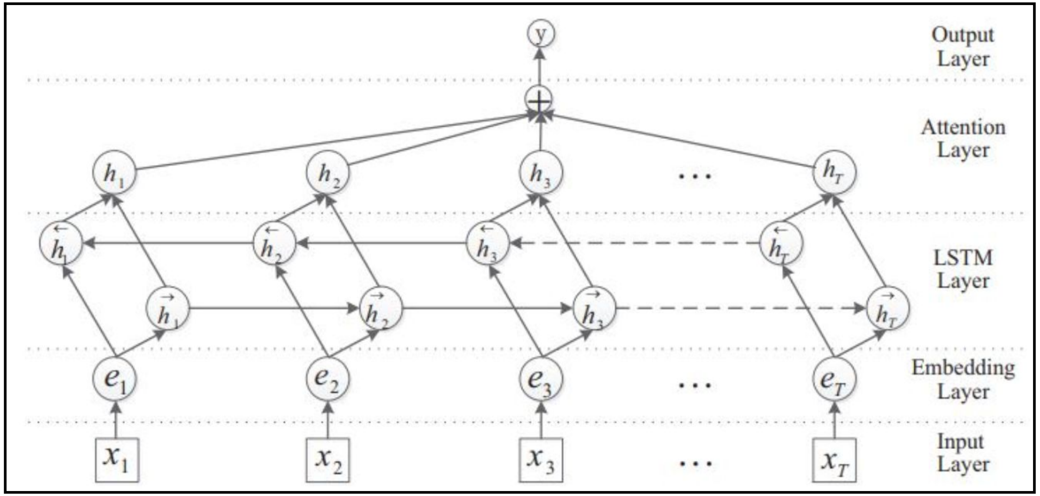
- Results: Embeddings not too helpful at separating the data.
    - Inherent structure of multi-topic articles theses days





Sentiment Analysis & Topic Classification

- <u>Idea:</u> Still can capture sentiment using embeddings.
  - Focus on classic Word2Vec to model corpus
  - Run through Bidirectional LSTM w/ Attention Layer
  - Capture either positive (1) or negative (0) sentiment



Sentiment Analysis & Topic Classification

The following is the classification score report on novel data:

	precision	recall	f1-score	support
0	0.57	0.61	0.59	12500
1	0.58	0.54	0.56	12500
accuracy			0.57	25000
macro avg	0.57	0.57	0.57	25000
weighted avg	0.57	0.57	0.57	25000

Unsupervised Sentiment Analysis

- Idea: Use topic modeling instead of explicit race mention for unsupervised sentiment analysis.
    - Apply Doc2Vec model on corpus
    - Use the similarity score to determine how well an article embodies the topic
    - Apply VADER and TextBlob on the articles to determine sentiment

Topic Keywords (Black): “black”, “african american”, “african-american”, “haitian”, “jamaican”, “west indian”, “dominican”

Topic Keywords (White): “white”, “irish”, “italian”, “caucasian”

Unsupervised Sentiment Analysis Results

VADER

Score Type	Black Articles	White Articles
Positive	0.0567	0.0591
Negative	0.0835	0.0822
Neutral	0.8598	0.8586

TextBlob

Score Type	Black Articles	White Articles
Positive	0.1168	0.1164
Negative	-0.0663	-0.0622

Entity Recognition:

- It is the task of identifying and categorizing key information (entities) in text, every detected entity is classified into a predetermined category.

What was accomplished:

- Entity recognition to match individuals in Data set
- Create Entity Categories that cover most of word in Data set
- Detect Entity name: accuracy is about 95%
- Detect Entity Street Name/Organization: accuracy around 93.5%

Entity Recognition Result:

```
Entity_Rec('wgbh_data_2018.csv')
```

```
Frank McClelland PERSON
Moncef Meddeb PERSON
McClelland. PERSON
Julia Child PERSON
Henry Kissinger PERSON
Mick Jagger PERSON
David Ortiz PERSON
Bill Belichick PERSON
McLelland PERSON
McClelland. PERSON
Judie PERSON
Amanda Beland PERSON
Jon Meacham PERSON
Lawrence O’Donnell PERSON
Caitlin Moran PERSON
Tom Papa PERSON
Richard Blanco PERSON
Harvard Historian PERSON
Nancy Koehn PERSON
```

Limitations / Possible Biases

- For crime coverage, not all incidents (namely, many street intersections) in the CSV could be geocoded with the current implementation
- Some error still exist in the entity recognition, for example:
  - WGBH maybe recognize as the person name
  - Elliot S! Maggin maybe recognize as “Elliot” and “Maggin”
- Inherent nature of news articles leads to subject embedding overlap
- Possibly existing confirmation bias as well

Future Work/Improvements:

- Continue to make improvements on how to better visualize the topic modeling results
- Link the person with the their organization after entity recognition
- Make use of a more robust geocoder or data that already has geographical coordinates for incidents of gun violence/homicide to analyze their coverage