Name: Sunay Sanghani USC ID: 5373536322
DSCI 510 Final Project

Abstract of the Project:

NewsPredictor is a project which concatenates data from disparate sources of news headlines, both local and national. The data of cleaned news headlines is then fed into the open-source large language Python library called nltk. Nltk's VADER framework is used to measure the sentiment analysis of the local and national headlines. The training set for the nltk sentiment analysis takes the large Kaggle dataset of headlines from Huffington Post. The Kaggle Dataset is a popular and robust data file which is commonly used as a base training set for natural language processing models. The test data set, which is what the sentiment analysis is measured on is the local and national headlines.

Findings from the sentiment analysis indicated that both national and local news headlines have a decent amount of negative sentiment headlines compared to positive sentiment headlines. Also, larger publications such as Huffington Post have a better chance of being more neutral and positive learning in their sentiment. There is more potential for analysis and findings based on geographic regions as indicated in my analysis. Geographic locations differ significantly in the sentiment of their headlines according to the sentiment analysis done in my project.

Motivation of the Project:

In today's interconnected world, it is extremely tough to follow news websites that are free of bias, negativity, and perceptions. Post pandemic, it also seems like there is more negativity overall in the sentiment of news channels with extremely polarizing headlines. However, I think that local news is more positive than national news sites. In this project, I want to test my hypothesis: the sentiment of local new headlines is more positive than the sentiment of national news websites.

The data will be collected through three sources: 1) New York Times API, 2) Web Scraping from local news websites, and 3) CSV of News headlines from Huffington Post. I plan to perform basic EDA on all datasets. San Jose Mercury News, Portland Observer, Chicago Tribune, Miami Sun-Sentinel, Boston Herald, and Denver Post are the local newspapers which I will use to scrape news headlines from. I tried to choose newspapers from different regions.

To calculate the sentiment analysis, I will use nltk, a built-in Python program which will analyze the headlines I scrape from a New York Times API(national) and pull from websites(local). It will use the Huffington Post headlines to train the model. Additionally, I plan to use matplotlib and other Python modules we learned in class to create visualizations that compare the sentiment of the local vs. national news. Additionally, I want to create geographic comparisons as well to maybe see if I can make conclusions about sentiment positivity of news based on geographic US regions. For example, are the southern US regions more positive in their news or the northeast regions?

Data Sets:

Data Source 1(News_Category_Dataset_v3.json): The CSV of Huffington Post news headlines will be downloaded as a CSV from Kaggle manually. I trust this dataset because it has been curated by a data scientist as a benchmark dataset for linguistic and statistical tests. The CSV is extremely large, with about 210K news headlines. I will use the dataset as a benchmark for statistical analysis for the local and national news as the author of the dataset mentioned it has relatively equal positive and negative sentiments. I will use Pandas to filter, clean, and keep relevant headlines useful for the training dataset. The robust dataset will make the sentiment analysis much more powerful and the model's accuracy better.
Example of entries in the JSON file:


```
{"link":
"https://www.huffpost.com/entry/covid-boosters-uptake-us_n_632d719ee4b087fae6feaac9", "headline": "Over 4 Million Americans Roll Up Sleeves For Omicron-Targeted COVID Boosters", "category": "U.S. NEWS", "short_description": "Health experts said it is too early to predict whether demand would match up with the 171 million doses of the new boosters the U.S. ordered for the fall.", "authors": "Carla K. Johnson, AP", "date": "2022-09-23"}

{"link":
"https://www.huffpost.com/entry/american-airlines-passenger-banned-flight-attendant-punch-justice-department_n_632e25d3e4b0e247890329fe", "headline": "American Airlines Flyer Charged, Banned For Life After Punching Flight Attendant On Video", "category": "U.S. NEWS", "short_description": "He was subdued by passengers and crew when he fled to the back of the aircraft after the confrontation, according to the U.S. attorney's office in Los Angeles.", "authors": "Mary Papenfuss", "date": "2022-09-23"}

{"link":
"https://www.huffpost.com/entry/funniest-tweets-cats-dogs-september-17-23_n_632de332e4b0695c1d81dc02", "headline": "23 Of The Funniest Tweets About Cats And Dogs This Week (Sept. 17-23)", "category": "COMEDY", "short_description": "\"Until you have a dog you don't understand what could be eaten.\"", "authors": "Elyse Wanshel", "date": "2022-09-23"}
```

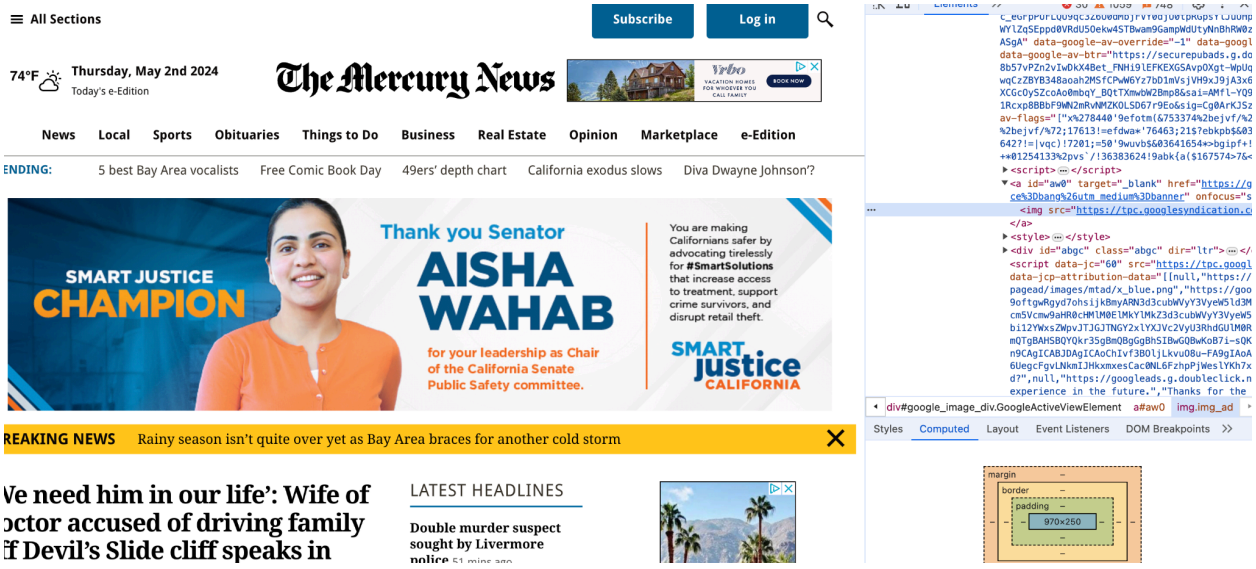

Data Source 2 (NYT_headlines.txt): Originally, I had planned on using the Tweepy API to scrape Tweets from local news Twitter accounts in my proposal. However, after tinkering with the API I realized that once I wanted to make more calls, it had a rate limit. The rate limit would require me to pay $1000 a month to access Twitter's v2 authorization(this rule changed recently after Twitter went private). When I implemented the method and tried to access the Tweets from the individual projects, it gave me an error asking for v2 authentication. The error is below:

```
Failed to post tweet: 403, {"errors":[{"message":"You currently have access to Twitter API v2 endpoints and limited v1.1 endpoints
```

only. If you need access to this endpoint, you may need a different access level.

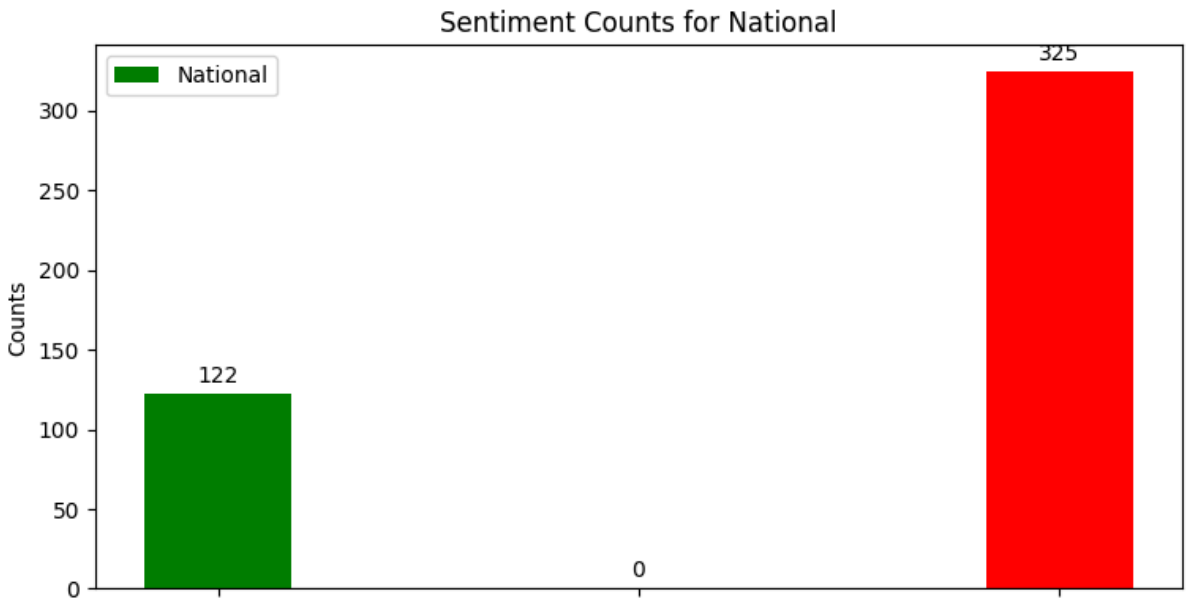
As a result, I was forced to change the plan and use the New York Times API instead of web scraping the New York Times website. The New York Times API has a bit of easier restrictions on the rate limits and it is easier to use it to access the headlines. The New York Times API is a suite of services offered by The New York Times that allows developers to access a variety of data from the newspaper in a programmatic way. The API provides multiple endpoints, each designed for different types of news articles and sections of the New York Times.

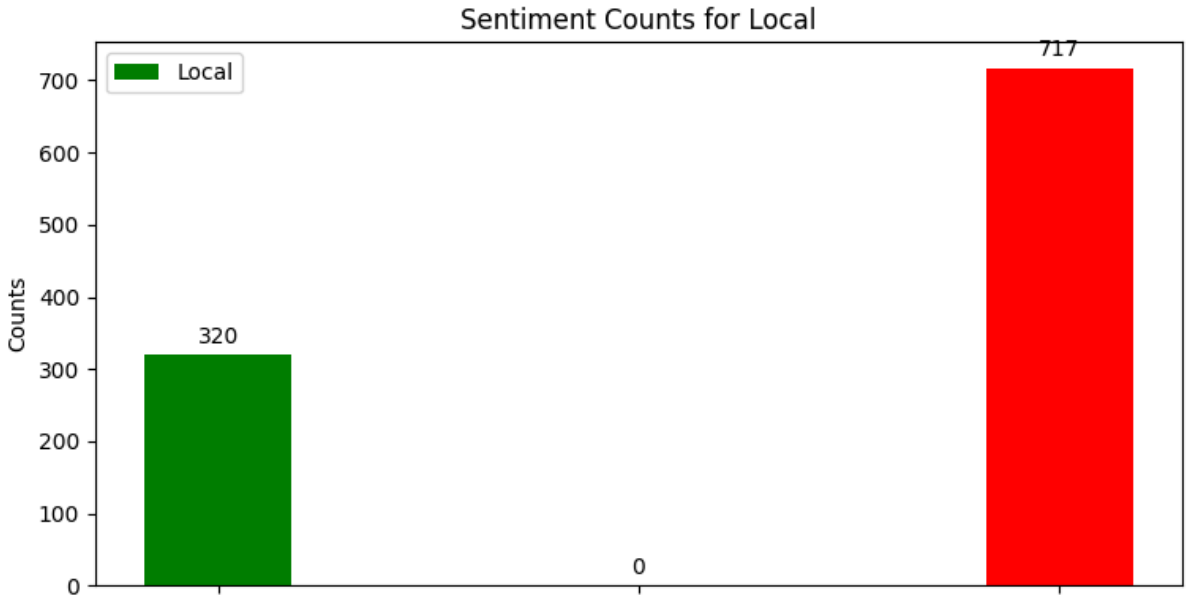
Data Source 3 (dataset_1.txt): The third data source is collected by scraping headlines from a news website from local newspapers. I was collecting only the headlines from recently published articles which I sorted by date. This dataset is particularly useful for projects involving trend analysis or sentiment analysis within particular geographic regions. Each website had a similar structure but different elements and different page layouts which required me to be cognizant of the fact that there are different HTML structures for these news sources.

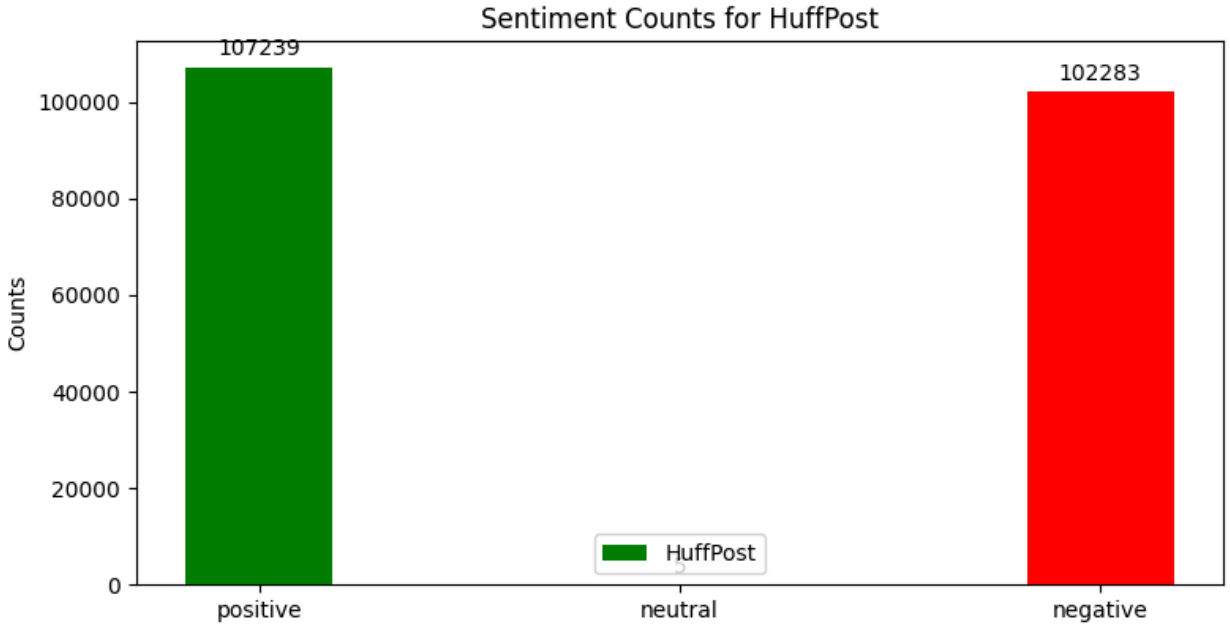


Analysis, Insights, Conclusions of Data and Graphs:

Bar Graph of Total Count

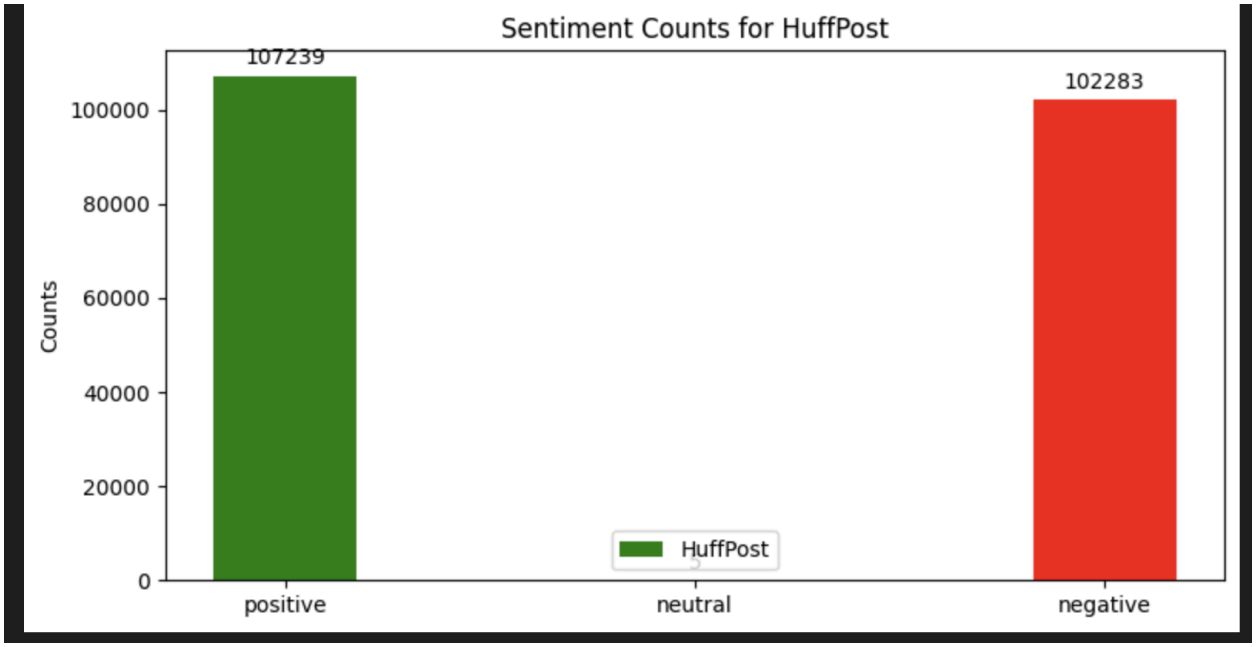

Sentiment Counts for National
National
Counts
325
122
0
300
250
200
150
100
50
0


Sentiment Counts for Local
Local
Counts
717
320
0
700
600
500
400
300
200
100
0


Sentiment Counts for HuffPost
Counts
107239
102283
HuffPost
100000
80000
60000
40000
20000
0
positive
neutral
negative

The National news shows a significantly higher number of negative sentiments compared to positive sentiments. The absence of neutral sentiments suggests that the national news items tend to evoke stronger opinions or reactions, possibly due to the nature of topics covered, such as politics, global events, or national crises.

Similar to the National news, local news also exhibits a higher count of negative sentiments over positive sentiments. The ratio of negative to positive sentiments is pretty similar to that of National news, indicating that local news, while focused on regional or local issues, may also cover many challenging or controversial topics that lead to negative sentiments. The absence of neutral sentiments underscores a trend toward more emotionally charged reporting or perhaps more impactful local events.



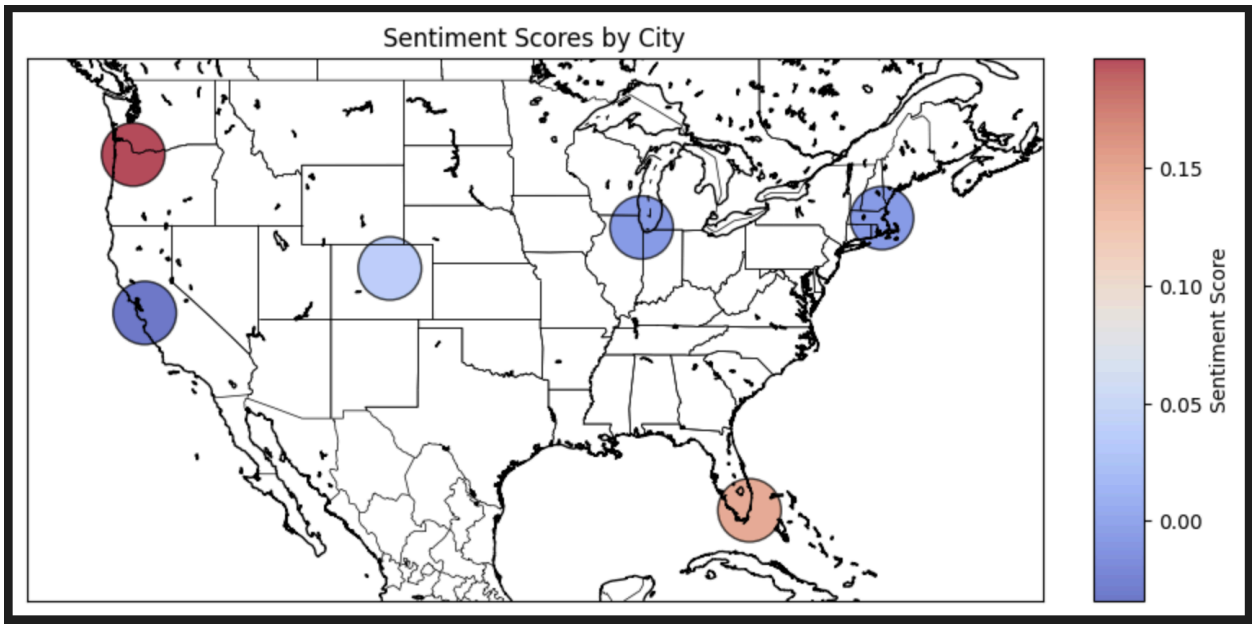
The sentiment distribution for the Huffington Post is more balanced between positive and negative sentiments compared to the National and Local news. However, it still leans slightly towards more positive outcomes. The presence of very few neutral sentiments is consistent with the other sources, indicating a general tendency for news content to be polarizing or at least clear in its emotional direction. The massive scale of counts reflects the broad coverage and large volume of content published by the Huffington Post.

Overall Takeaways:

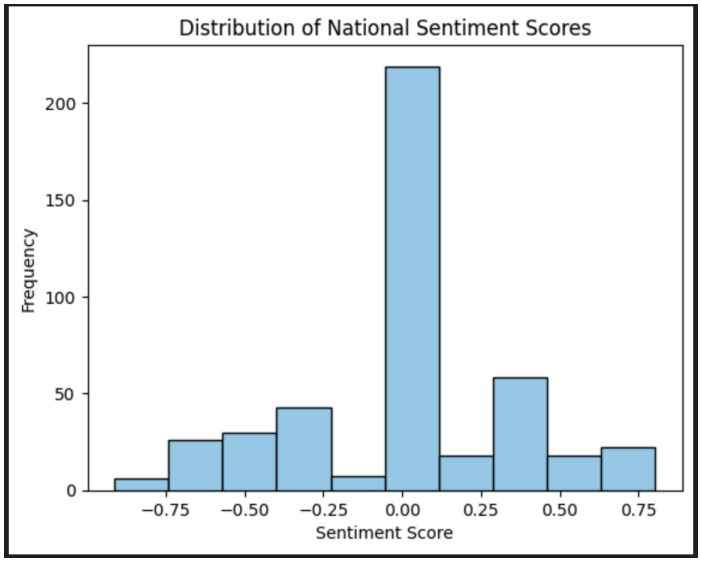
- Negativity Dominance: All three sources show a higher number of negative sentiments compared to positive ones, which is typical in news media due to the greater impact of negative news on readership and engagement.
- Impact of News Type: National and Local news, while different in scope, show similar trends in sentiment distribution, hinting at the universally challenging nature of news topics regardless of scale.

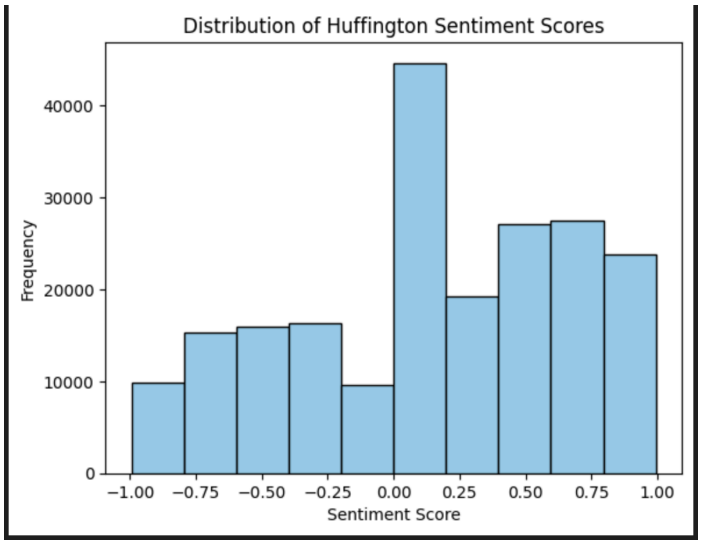
Geographic Sentiment Analysis

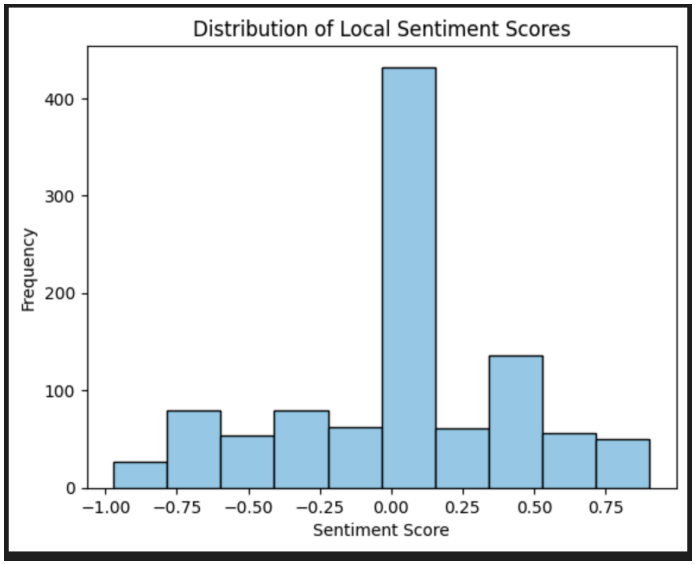
After examining each source, I curated the data to perform sentiment analysis on the headlines of the local newspapers. I had data files created by location(labeled in zip file as {location}_headlines.txt) as well so I plugged those into the sentiment analysis tool and used geopandas and matplotlib to create a map of sentiment analysis scores.



The figure above aims to measure variability in the sentiment among different sentiment scores in the city. There is Geographic Sentiment Variability in the US. For example, the San Jose Mercury News source has an extremely low average sentiment score with all of the articles. In contrast, the southern region has a deep red color which means it has a higher average sentiment score. Almost all cities and geographic regions have a different sentiment score along the region. I find this geographical sentiment analysis to be fascinating as it could have multiple implications for business or political insights. For example, the map hints at regional differences in public mood or media tone, which could be useful for businesses, political campaigns, or social researchers trying to understand regional dynamics.


Distribution of National Sentiment Scores
Frequency
Sentiment Score
0
50
100
150
200
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75


Distribution of Huffington Sentiment Scores
Frequency
Sentiment Score
0
10000
20000
30000
40000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Distribution of Local Sentiment Scores
Frequency
Sentiment Score
0
100
200
300
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75

I wanted to check the average sentiment scores of the headlines to check the spread of the sentiment scores. The bar graphs above examine the spread of the sentiment scores for all headlines in each dataset. The overall sentiment and spread is about the same in each of the bin distributions. Most of the article headlines are at a neutral to slightly positive levels. There was not any significant difference amongst the spread.

A big note for all the analysis above and my project in general is that the headline data is time sensitive. The headlines being pulled in my project vary from day to day. Some days headlines could be positive and other days they could be negative. As a result, the analysis I am doing on this day(May 2) could vary severely from other days and the graphs could also look different.

Conclusion

My hypothesis was incorrect. Local news headlines are not more positive than the sentiment of the national news headlines. Even for larger, more global publications like the Huffington Post, they have a slightly balanced sentiment analysis compared to local and national news sources. The data analysis also indicates that there are geographical differences in the sentiment as pertaining to individual regions. My project shows that there are sentiment analysis differences between each region and some of the differences are stark which gives insights into American journalism.

Limitations, Challenges, Future Work Ideas

With more time and scope, I would have liked to collate more data sources from different local news sources. I believe that the most interesting analysis with most insights would be the local news headlines sentiment analysis disparity. Each geographic location in the US has such a stark contrast in political, religious, and urban viewpoints that it is important to analyze those metrics. In the future, I would like to expand upon the data source and scrape more news sites. In terms of challenges, it was tough to navigate the different APIs. There are many APIs that offer good news headline scraping tools however the rate limits are pretty unforgiving. The rate limits that exist only offer 100 requests/day which makes it difficult to get a robust data set. Other limitations I had were navigating the different format of the HTML on the websites for the web scraping. Overall, it was a great learning experience to manage and clean different data sources.