Analyzing Social Trends on Recession conditions

Kanishk Barhanpurkar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
kbarhan1@binghamton.edu

Harshad Bhandwaldar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
hbhandw1@binghamton.edu

Nikita Mandlik
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
nmandli1@binghamton.edu

Brinda Eshwar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
beshwar1@binghamton.edu

ABSTRACT

Social Media platforms are widely used for transmitting data in different formats. We can generate a lot of information on recent trends. Additionally, public opinions can be used as a tool to predict forthcoming events. We are collecting data from two social networking sites, Twitter and Reddit. We are also using News Articles as another data source. Twitter is a public social networking domain and Reddit is a community-based social media platform. The primary aim of this project is to analyze the social media-associated data which predicts the conditions which lead to the Recession over time. Secondly, for the third dataset, we evaluated the news article which contains the topic of recession and economic crisis. techniques to understand the data in a better context. We have used data visualization techniques based on sentiment analysis, keyword analysis, and word length. The parameters help to understand the validity and initial insights of the data.

1. INTRODUCTION

A recession is a business cycle contraction when there is a general decline in economic activity. In 2020, Covid-19 spreads exponentially in the entire world due to which many nations imposed a lockdown on the entire nation. This decision affects international trade and decreases the currency flow among various countries [1]. Several countries whose economy is dependent on the tourism sector were affected badly. Therefore, the major output of this can be seen in 2022 which results in inflation and a lack of centralized money to fulfill citizens' basic needs. Social-networking sites like Twitter and Reddit play very vital roles in understanding public opinions [2]. The number of users using social networking sites makes it mandatory to analyze the trend of current affairs happening across the globe. In this project, we are using Twitter Data (Twitter API), Reddit Data (Reddit API) as public-opinion data, and NYTimes data as a news article dataset for recession conditions [3] [4] [5]. We are fetching the ID, text, and timestamp from Twitter while collecting the id, current DateTime, and postdate from the analysis. In the New York Times API, we are using the abstract summary, headline, publication timeline (date, time, and year), and word count. We have pre-defined the topic of the Recession and added a few keywords as filters to the data-scraping script. Additionally, we will be filtering the tweets and Reddit posts data based on the semantics of English languages. In this project phase, the main objective will be to collect the optimum amount of data, perform data cleaning and preprocessing and perform data analysis. We will be using different data visualization techniques for understanding the data better.

For the data analysis, we have used the keyword analysis from the data set. A keyword analysis is a technique to generate the most recurring words and important words present in the dataset. We have used the WordCloud library based on Python to perform the keyword analysis [6]. We have also performed the word-frequency analysis where we have analyzed the number of words that appeared in each tweet, subreddit comment, and NYTimes post. We have used the Seaborn and Matplotlib library (Python environment) for the data analysis [7][8]. We have also performed sentiment analysis for each tweet (Twitter), each comment of subreddit data (Reddit), and each article of New York Times data. We have used the Textblob library for sentimental analysis and used matplotlib to visualize the data in form of a histogram [9].

2. RESEARCH QUESTIONS

The primary objective of the project is to describe the influence of the recession and economic conditions based on public opinions and newspaper data. The research question for the project is as follows-

a. How do public opinion and sources of information (newspapers) associated with each other for recession topics?
b. How does the sentimental analysis score change for the recession conditions over time?
c. How do social media platforms influence the recession conditions?

3. RELATED WORK

Twitter data is analyzed within the context of post-pandemic effects on economic factors. Additionally, the statistical analysis provides an in-depth insight into the validity of the data collection method (Rahman M. et al., 2020). The Twitter data also helped many multinational companies to acquire customer knowledge on social media analysis (He, W. et al., 2018). In this research study, the authors have described the significance of social media analytics (Reddit) to understand the basic needs of citizens during the pandemic era (Lee, J. Y. et al, 2020). New York Times data has been analyzed from 1981 to 2021 and found that public opinion can help in detecting the recession conditions (Schmitt, A. J., 2020).

4. RESULT DISCUSSION

We are comparing the data that we have collected from the tweets, Reddit posts, and the articles from NYTimes using four parameters.

4.1 Time Series Analysis

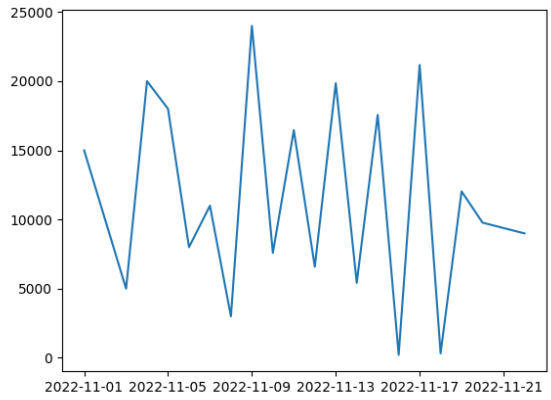

Figure 1: Graph representing data collected over time from recession-related tweets.

Figure 1 consists of a graph illustrating the data collected from tweets related to the recession over the span of 21 days starting from November 1st to November 21st. The x-axis of the graph represents the dates when the data is being collected, and the y-axis represents the number of tweets we have collected on particular dates. We have collected 155,218 tweets collected using Twitter API.

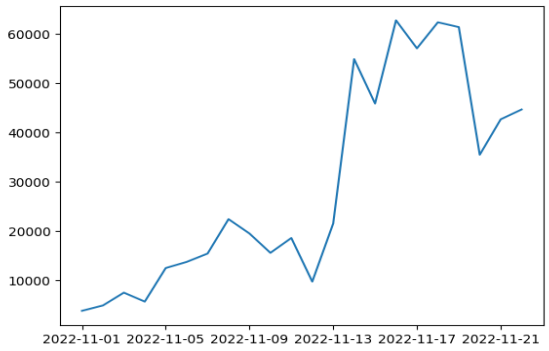

Figure 2: Graph representing data collected over time from recession-related Reddit posts.

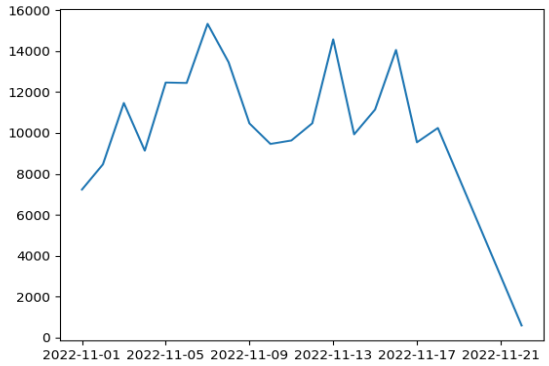

Figure 3: Graph representing data collected over time from recession-related NYTimes articles.

Figure 2 consists of a graph illustrating the data collected from Reddit posts related to the recession over the span of 21 days. The x-axis of the graph represents the dates when the data is being collected, and the y-axis represents the number of Reddit posts we have collected on particular dates. Figure 3 consists of a graph illustrating the data collected from NYTimes articles related to the recession over the span of 21 days. We have collected 240,079 articles and 616,895 Reddit comments from the NYTimes and Reddit API respectively.

S.No.	Dataset	Number of records	Memory-usage (in MB)
1.	Twitter	155,218	40.58
2.	NYTimes	240,079	49.58
3.	Reddit (r/Recession)	616,895	62.59
4.	Reddit (r/Politics)	1,213,499	132.67

Table 1: Comparison of Number of Records and Memory usage for various datasets.

4.2 Word-cloud analysis (keyword)

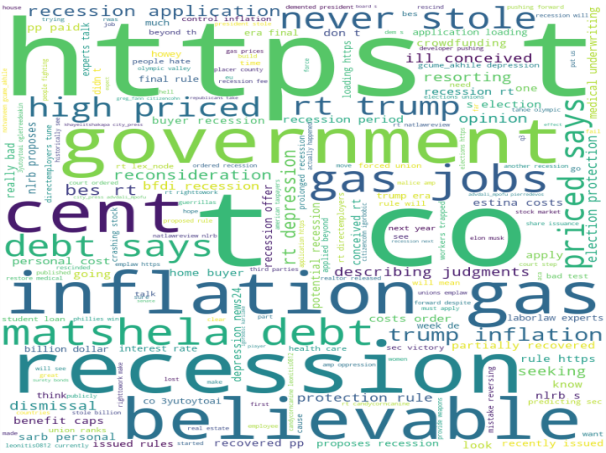

Figure 4: Word-cloud for the Twitter data

Figure 4 consists of a word-cloud for the data collected from the Twitter data. The words that are commonly used represent the words related to the recession. The words like ‘government’, ‘inflation’, ‘gas’, ‘recession’, and ‘priced’ are the most common words used in tweets by users.

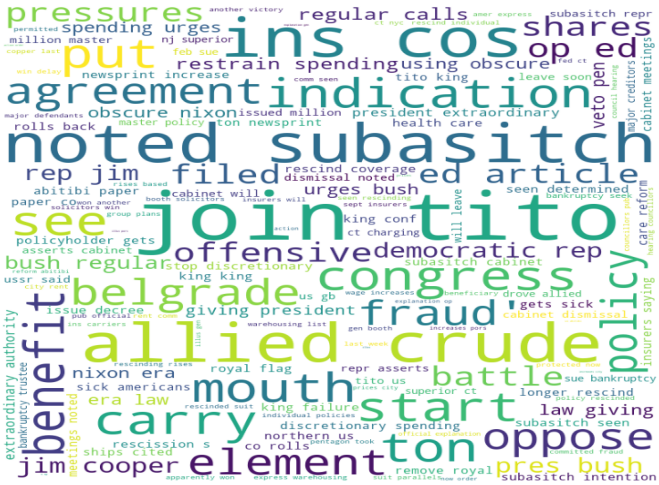

Figure 5: Word-cloud for the New York Times data.

Figure 5 consists of the word-cloud that is being created based on the data that was collected from the NYTimes articles. The keywords like ‘agreement, ‘oppose’, ‘allied’, ‘crude’ and ‘indication’, and others are most commonly used in articles related to the recession. Figure 6 consists of the word-cloud that is being created from the data collected from the Reddit posts that were related to the recession. The words like the ‘stock-market’, ‘worst decline’, ‘recession’, ‘financial crisis’ and ‘consumer’, and others are the most commonly used in the posts for r/Recession on Reddit.

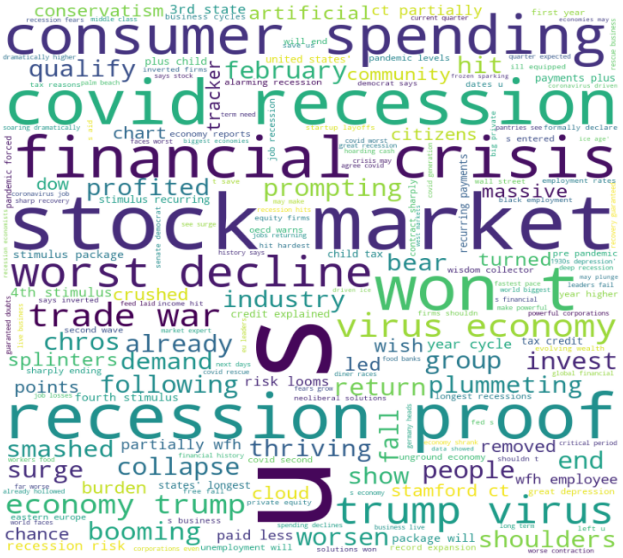

Figure 6: Word-cloud for the Reddit data.

In Figures 4 and 6, the keywords are very common as the results are based on public-opinionated datasets (Twitter & Reddit) whereas, in Figure 5, the keywords are more generalized.

4.3 Memory Usage analysis

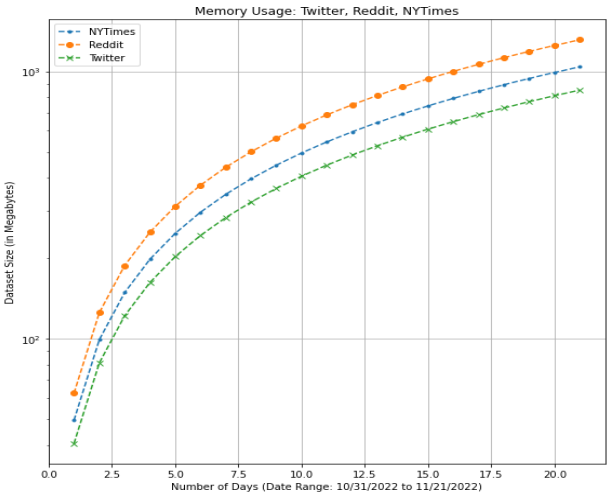

Figure 7: Memory-usage of each data source per day

Figure 7 consists of a graph in which the x-axis represents the number of days we collect the data from the data sources. The y-axis represents the dataset size which is the size of data collected for the corresponding days. This graph provides an insight into the Memory usage done by each dataset as the data sources like Reddit, Twitter, and NYTimes produce very different kinds of data.

4.4 Word Frequency analysis

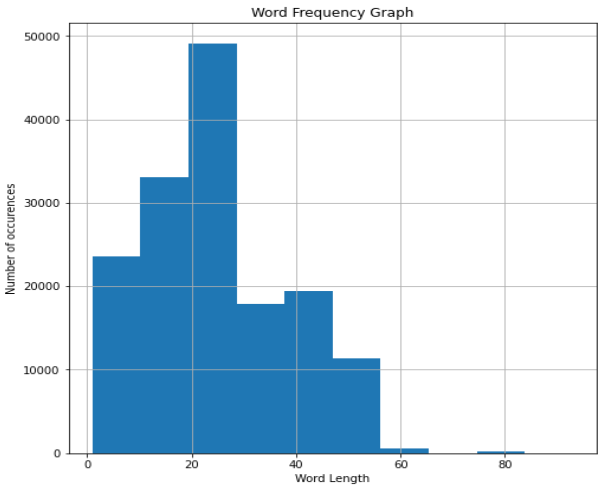

Figure 8: Frequency graph for the number of words for every tweet.

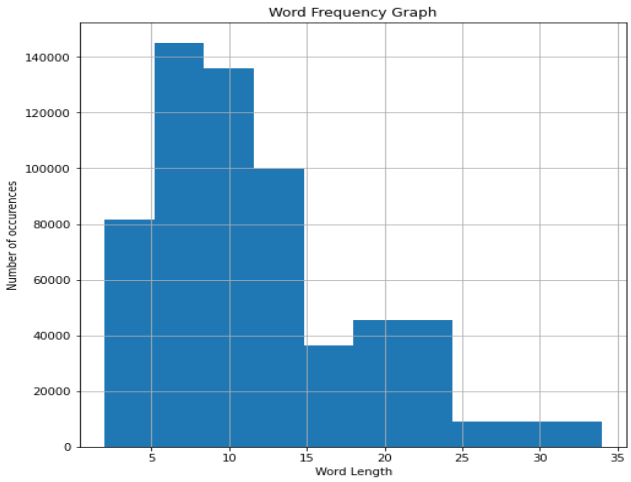

Figure 9: Frequency graph for the number of words for every Reddit comment.

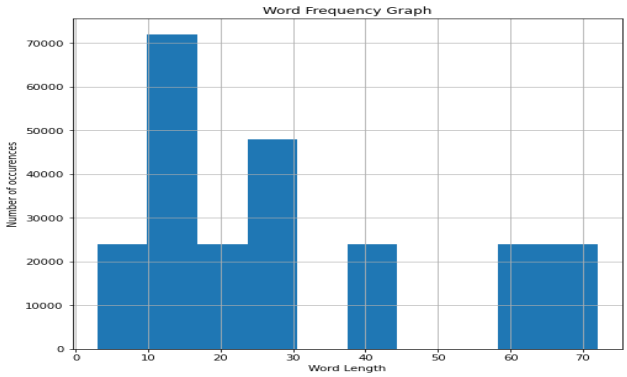

Figure 10: Frequency graph for the number of words for every NYTimes article (abstract).

Figures 8, 9, and 10 consist of a graph representation in a tweet (Twitter), Reddit comment (Reddit API), and NYTimes article abstract respectively where the x-axis represents the number of words in each particular record and the y-axis represents the number of occurrences. In the Twitter data, maximum occurrences of words can be obtained in the range of 20-40. The Reddit data contain the maximum number of occurrences in the range of 5-15. Additionally, the NYTimes data collected in the abstract form of the article also contains a range between 10-30.

4.5 Sentiment score analysis

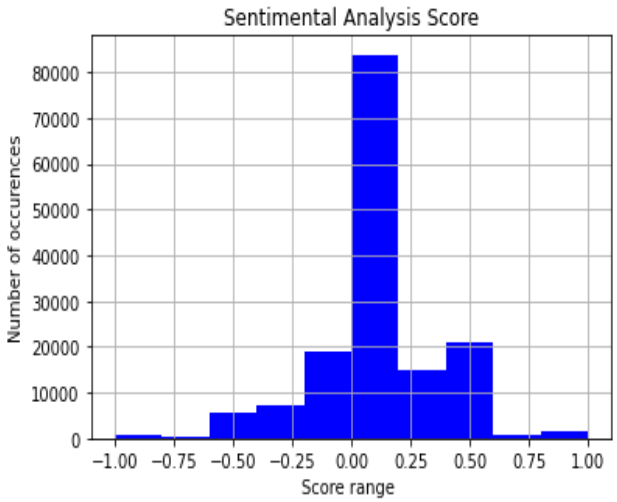

Figure 11: Sentimental analysis scores of the tweets and their occurrences

Figure 11 consists of a graph in which the x-axis represents the sentiment analysis score ranges of different tweets related to the recession. The score range 0.0 to 0.25 has the highest occurrence in this graph.

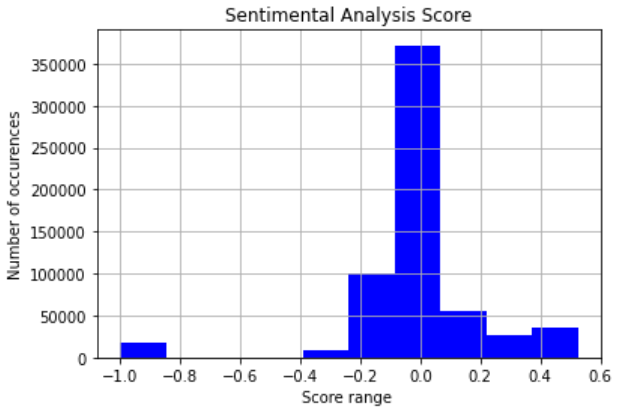

Figure 12: Sentimental analysis scores of the Reddit posts and their occurrences

Figure 12 consists of a graph in which the x-axis represents the sentiment analysis score ranges of different Reddit posts related to the recession. The score range -0.2 to 0.0 has the highest occurrence in this graph.

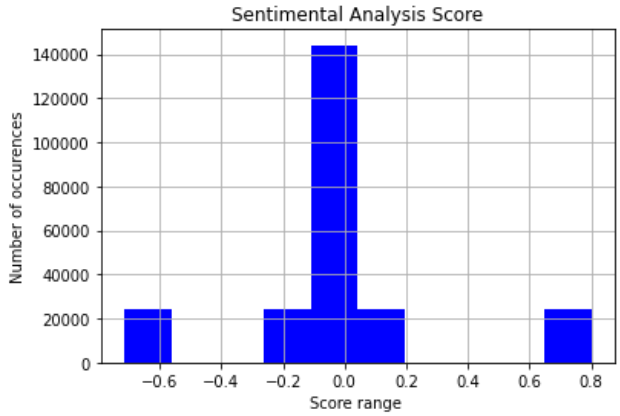

Figure 13: Sentimental analysis scores of the NYTimes Articles and their occurrences

Figure 13 consists of a graph in which the x-axis represents the sentiment analysis score ranges of different NYTimes Articles related to the recession. The score range -0.2 to 0.0 has the highest occurrence in this graph. In Figure 11, the Twitter data polarity is evenly distributed in which extreme positive and negative patterns are observed. The Reddit data contain moderately negative correlation and strong positive correlation. Additionally, we observed that NYTimes data shows properties of low positive, low negative, strong positive, and strong negative attributes.

5. Reddit data (r/Politics) sub-reddit analysis

5.1 Time-series analysis

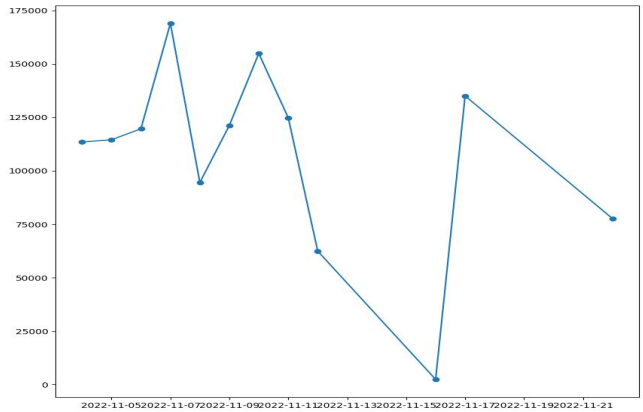

Figure 14: Graph representing time-series analysis.

Figure 14 consists of a graph illustrating the data collected from Reddit posts related to politics over the span of 21 days. The x-axis of the graph represents the dates when the data is being collected, and the y-axis represents the number of Reddit posts we have collected on particular dates.

5.2 Word-cloud analysis

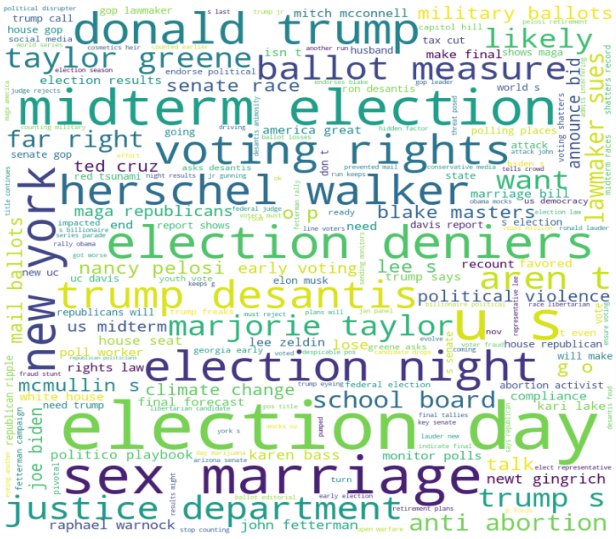

Figure 15: Word Cloud for r/Politics sub-reddit data.

Figure 15 consists of the word cloud that is being created from the data collected from the Reddit posts that were related to politics. The words election day, school board, and others are the most commonly used in the posts for r/politics on Reddit.

5.3 Frequency of characters per r/Politics

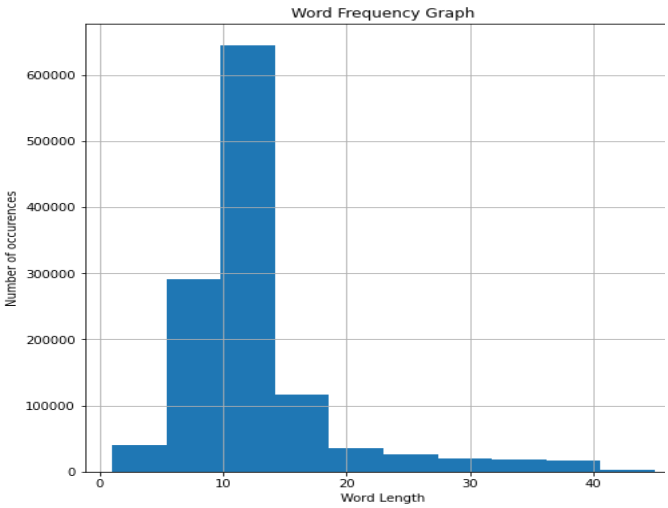

Figure 16: Frequency graph for the number of characters

Figure 16 consists of a bar graph in which the x-axis represents the number of characters in a r/politics post and the y-axis represents the number of occurrences.

5.4 Sentiment Score Analysis

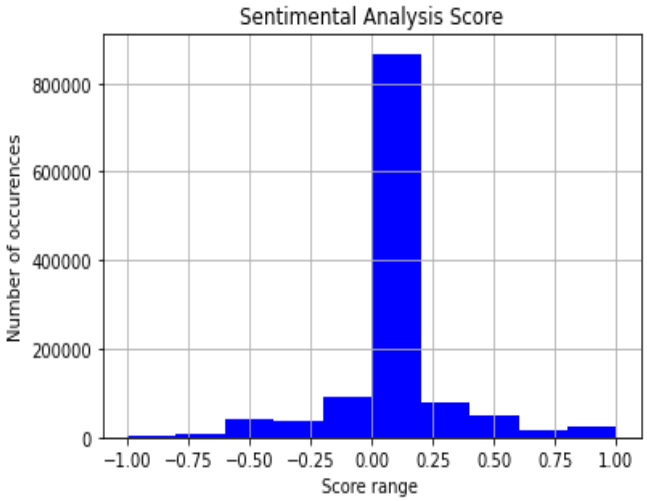

Figure 17: Sentimental analysis scores of the reddit data related to politics and their occurrences

Figure 17 consists of a graph in which the x-axis represents the sentiment analysis score ranges of different NYTimes Articles related to politics. As we can see the score range 0.0 to 0.25 has the highest occurrence in this graph.

6. CONCLUSION

We build the second stage of the pipeline from two social networking sites- Twitter and Reddit. Also, we have used the source of information medium i.e. NYTimes API, and compared the results for the public opinionated data and a source of information for the recession conditions. We have also analyzed the data based on keyword analysis, sentimental analysis score, word length comparison, and data collected for the given time interval. Additionally, we have also analyzed the r/Politics subreddit for the recently held midterm elections and analyzed the data based on the key factors mentioned above.

7. FUTURE SCOPE

After the initial analysis, we will be creating a web-based application using Python frameworks to visualize the data based on sentimental analysis. It will also provide insights about the result based on the public opinionated social networking sites and sources of information like NYTimes newspaper articles.

REFERENCES

[1] Feyisa, H. L. (2020). The World Economy at COVID-19 quarantine: contemporary review. International journal of economics, finance and management sciences, 8(2), 63-74

[2] Mariolis, T., Rodousakis, N., & Soklis, G. (2021). The COVID-19 multiplier effects of tourism on the Greek economy. Tourism economics, 27(8), 1848-1855.
[3] TwitterAPI Documentation.
https://developer.twitter.com/en/docs/twitter-api
[4] RedditAPI Documentation.
https://www.reddit.com/dev/api/
[5] NYTimes API Documentation.
https://developer.nytimes.com/docs/articlesearch-product/1/overview
[6] WordCloud Library Documentation.
https://pypi.org/project/wordcloud
[7] Seaborm Library Documentation.
https://seaborn.pydata.org/
[8] Matplotlib Library Documentation
https://matplotlib.org/
[9] Textblob Library Documentation.
https://textblob.readthedocs.io/en/dev/
[10]. Rahman, M., Ali, G. G., Li, X. J., Paul, K. C., & Chong, P. H. (2020). Twitter and census data analytics to explore socioeconomic factors for the post-covid-19 reopening sentiment. arXiv preprint arXiv:2007.00054.
[11]. He, W., Zhang, W., Tian, X., Tao, R., & Akula, V. (2018). Identifying customer knowledge on social media through data analytics. Journal of Enterprise Information Management.
[12]. Lee, J. Y., Chang, O. D., & Ammari, T. (2021). Using social media Reddit data to examine foster families' concerns and needs during COVID-19. Child abuse & neglect, 121, 105262.
[13]. Schmitt, A. J. (2021). Counting on the News: Data and Sentimentality. A 40-year Text Analysis of The New York Times.