Contents

1 Introduction

I will be looking at a variety of social network analysis and data collection strategies in this Report. The first step entails searching for current trends on Twitter, analyzing those tweets to determine what kind of devices they originated from. The second objective entails visualizing a social network graph, such the graph of the Amazon Network, determining some centrality metrics for said graph, and then carrying out community detection. The third step entails gathering tweets regarding a specific campaign or event, performing exploratory data analysis, and finally completing sentiment analysis. Collecting news articles, performing subject modelling, and then producing a written summary make up the fourth and final duty.

2 Twitter Trend Analysis

A networking site called Twitter allows users to create, share, and consume digital material. Users are always free to create and post a "tweet" on any topic they desire. A issue becomes a "trend" if lots of people are tweeting about it. When looking for breaking news before it is covered by mainstream media, Twitter trends may serve as a useful resource.

Twitter trends can relate to entertainment, sports, and cultural events in addition to news and current affairs. For instance, during important award ceremonies or sporting events, Twitter is bombarded with tweets about the event or game, making it simple to follow the most well-liked moments and comments in real-time.

2.1 Current Popular Trends in the UK

Using the get_place_trends API of Twitter Developer, we gathered the most popular subjects on Twitter as of April 29th, 2023. The resulting data was box plotted in Figure 1 to show the popularity of the various topics, with "Charles" receiving over 200,000 tweets as the most popular. According to F1 News in 2023, an incident that happened during the F1 2023 Azerbaijan Grand Prix Sprint was the cause of the spike in tweets.

The top three trending topics on that day, as shown in Table 1, were "Charles," "Leclerc," and "May Day." In Figure 2, these trends were further highlighted.

Overall, these patterns highlight the effectiveness of social media in delivering up-to-the-minute information and viewpoints on news and current events. Businesses, organizations, and people wishing to keep informed and involved with the most recent issues and debates may find it useful to follow and analyze Twitter trends in order to gain this information.

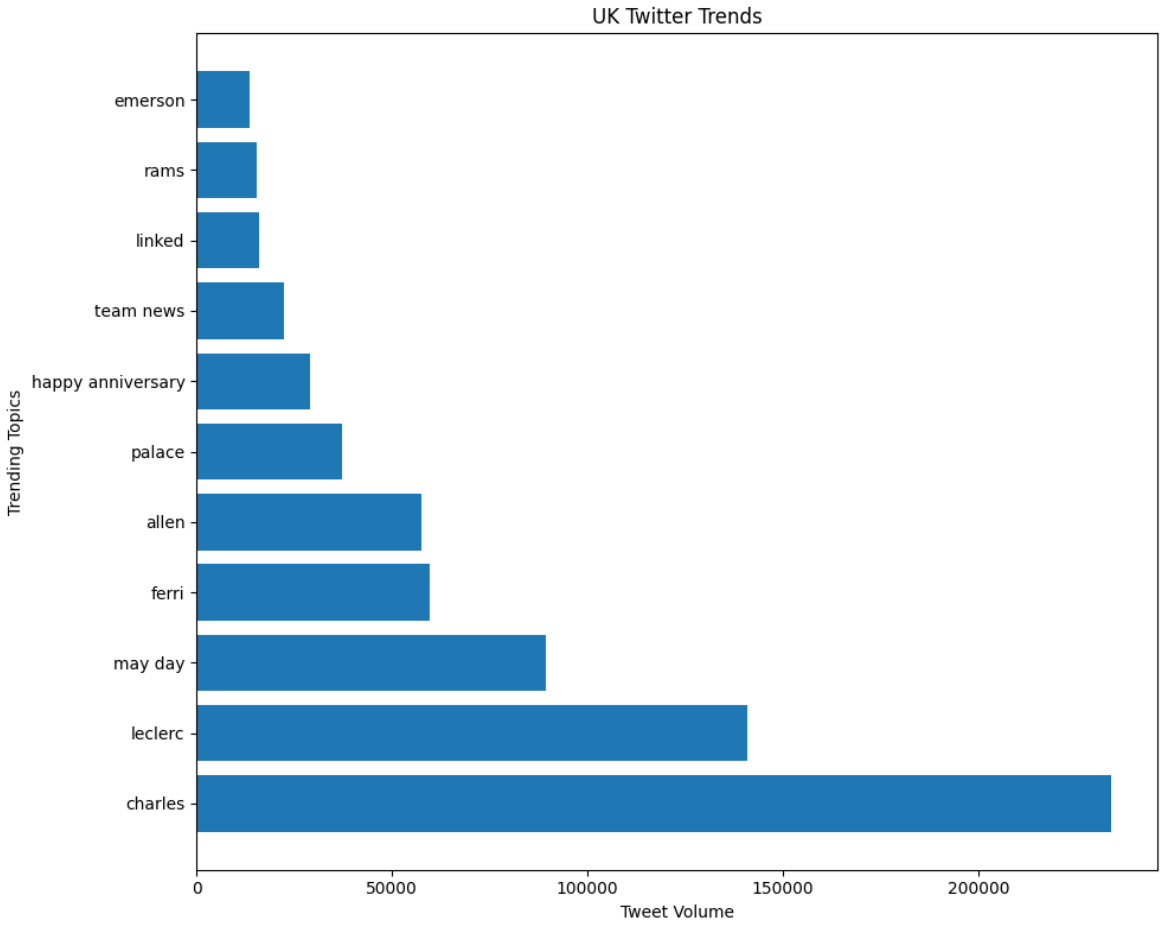


Figure 1: Current tweet trends in UK

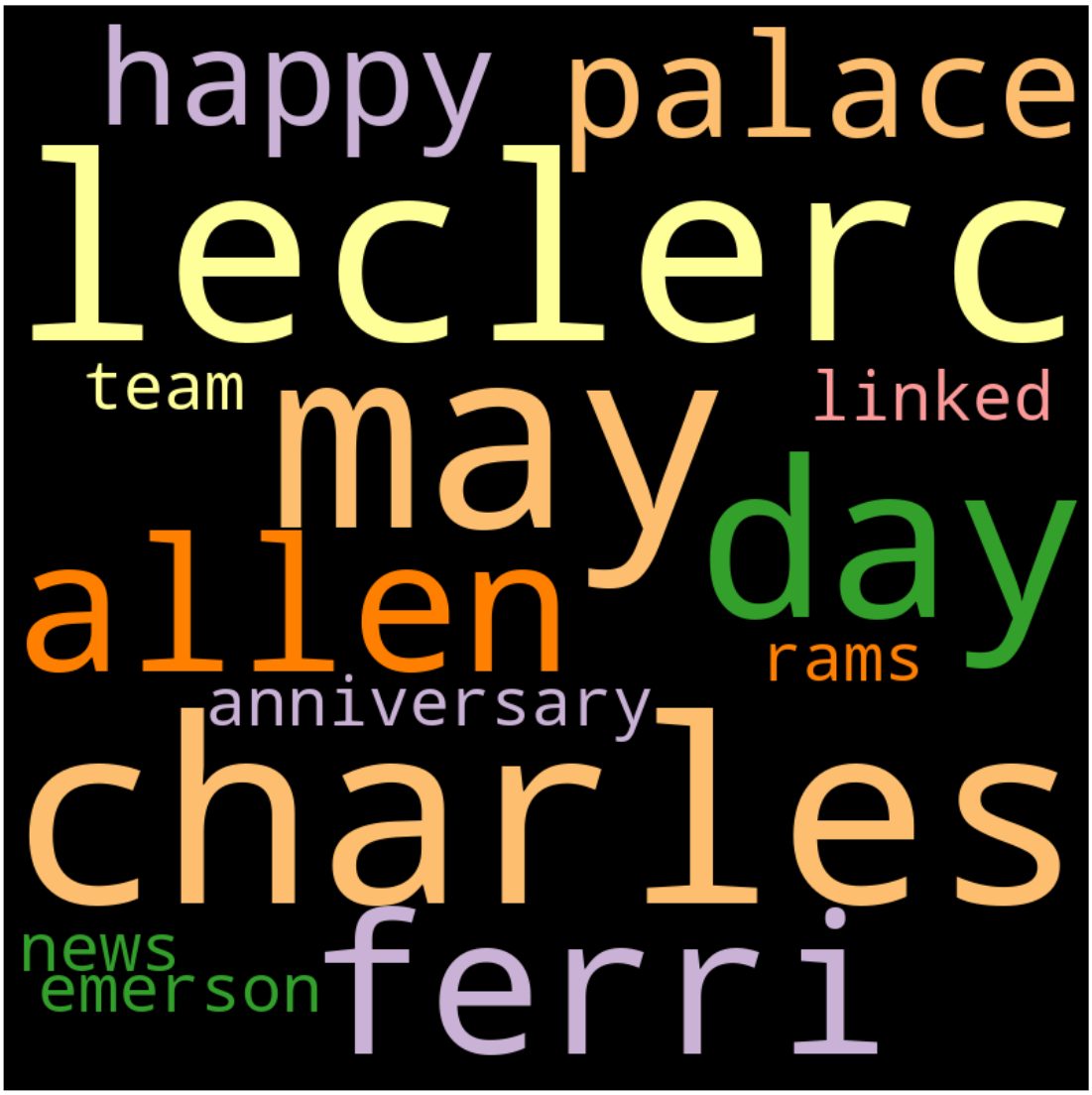


Figure 2: Current UK tweet trends word cloud 29/04/2023

Rank	Trend
1	Charles
2	Leclerc
3	May Day

Table 1: Top 3 trending topics of 29/04/2023

2.2 Tweet Sources

The sources of the collected tweets are shown in Figure 3, which reveals that the majority of users (51.8%) accessed Twitter using mobile devices running either the Android or iOS operating systems. The top three sources in Fig 3, which make up 87.3% of the overall tweet volume, further demonstrate this. The tweets collected are 10000 and the count of Tweets are mentioned in Table 2.These results underline the growing significance of mobile technologies in the social media sphere. Businesses and organizations must make sure their content is optimized for mobile viewing and engagement because the majority of Twitter users use the service through mobile devices. Businesses may better target their social media strategy to reach and interact with their target audience by studying the sources of tweets and user behaviour. These are the best insights that are found.

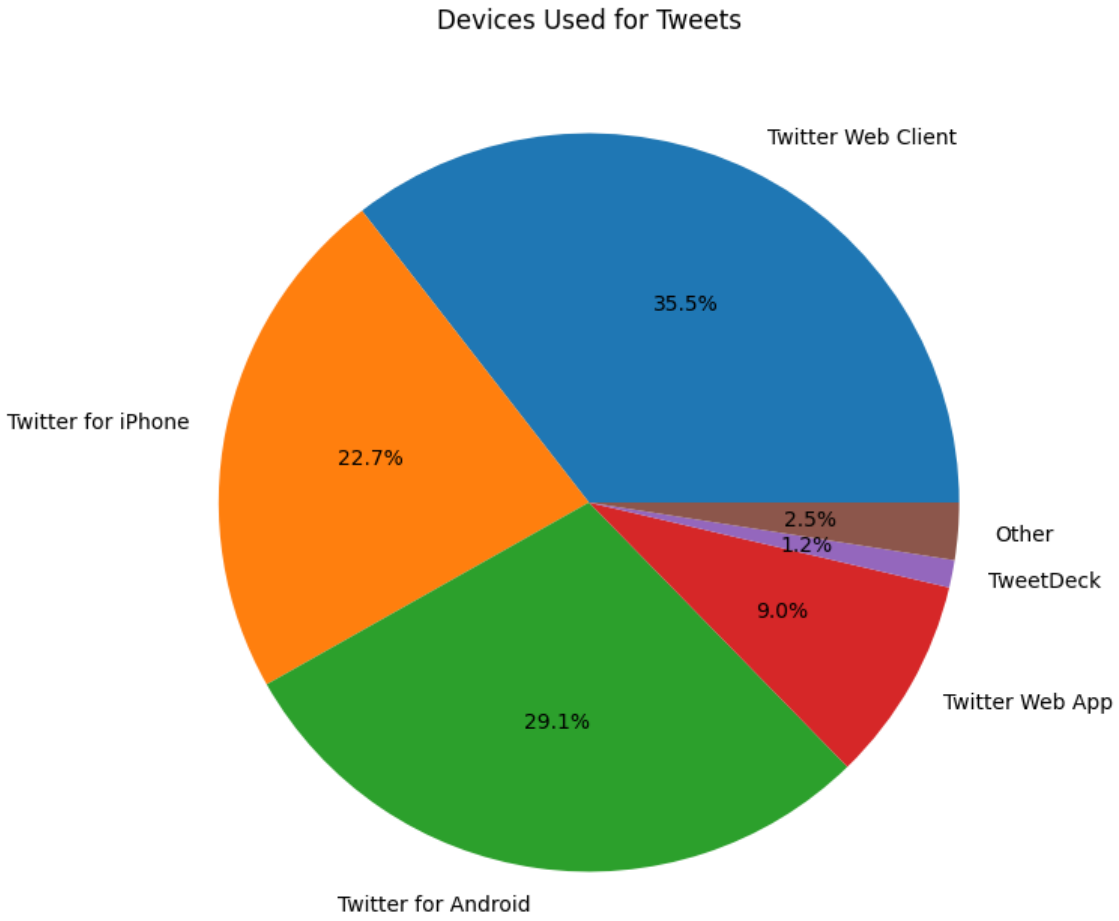


Figure 3: “Charles” trend tweet sources

Source	Quantity of Tweets
Twitter web Client	3550
Android Users	2910
iPhone Users	2270

Table 2: Top three tweet sources

2.3 Source Reliability

It might be difficult to determine the credibility of sources in tweets. Analyzing the source of the tweet is one approach that has been thought of because automated messages are less likely to be sent through official Twitter applications. Figure 3 shows the distribution of tweet sources and demonstrates that since all of the gathered tweets came from unofficial Twitter clients, they can all be regarded as being somewhat unreliable.

Whether or not the tweet's author is verified by Twitter was looked at as another indicator of credibility. No verified users were found among the 10000 tweets that were examined, proving that all of the tweets came from non-verified individuals.

These results underline the significance of exercising caution when analyzing and relying on data obtained from social media sources. When evaluating the reliability of information on social media platforms, it's critical to take a variety of factors into account, even though some reliability metrics can offer insightful information. There are no verified accounts in the data as of now due to removal of verified accounts by twitter for all notable personalities and making the verified account as a chargeable service.

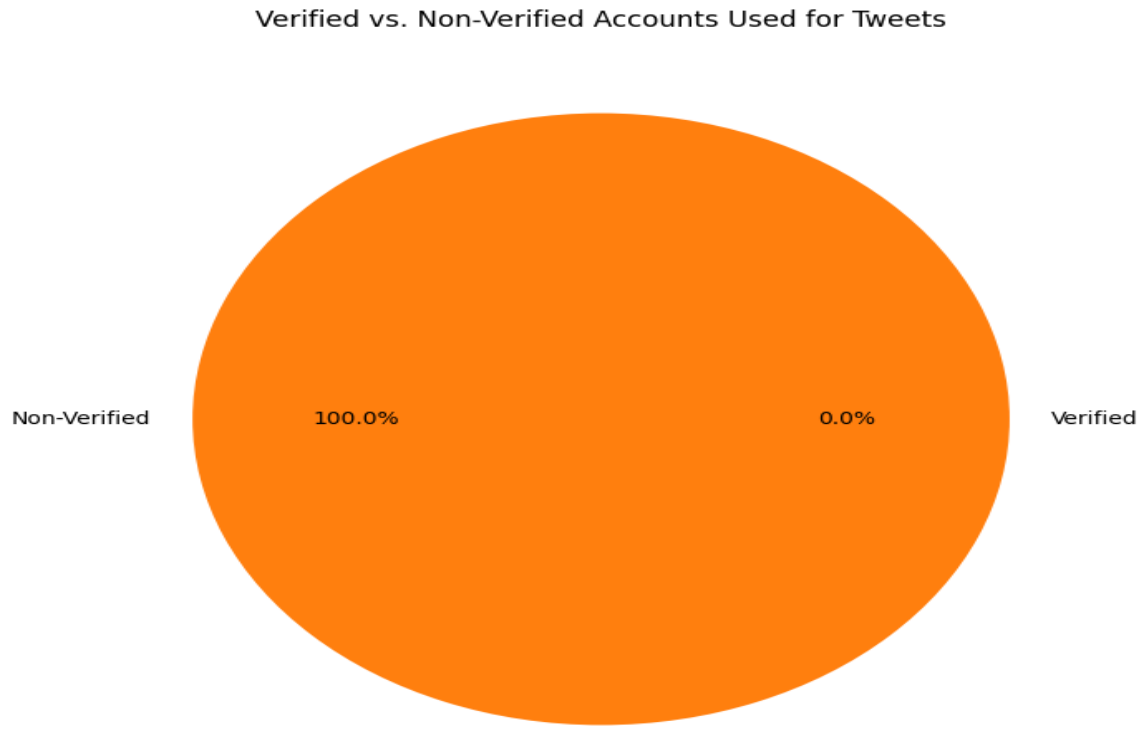


Figure 4: “Charles” trend verified user distribution.

3 Amazon Network Graph Analysis

A number of centrality measures, including degree centrality, proximity centrality, betweenness centrality, and eigenvector centrality, are computed for each node in this set of Amazon Network Data, which depicts a network of relationships between nodes (each node being represented by an ID).

The script loads a dataset from a CSV file that contains details about the graph's edges, uses NetworkX to generate a directed graph, and then determines the various centrality indices.

The top 10 nodes for each centrality metric are then written out as a result of the results. In addition, the script generates network visualizations based on degree, betweenness, and eigenvector centrality, prints the number of communities discovered by the Louvain method in the network, and displays a visualization of the communities.

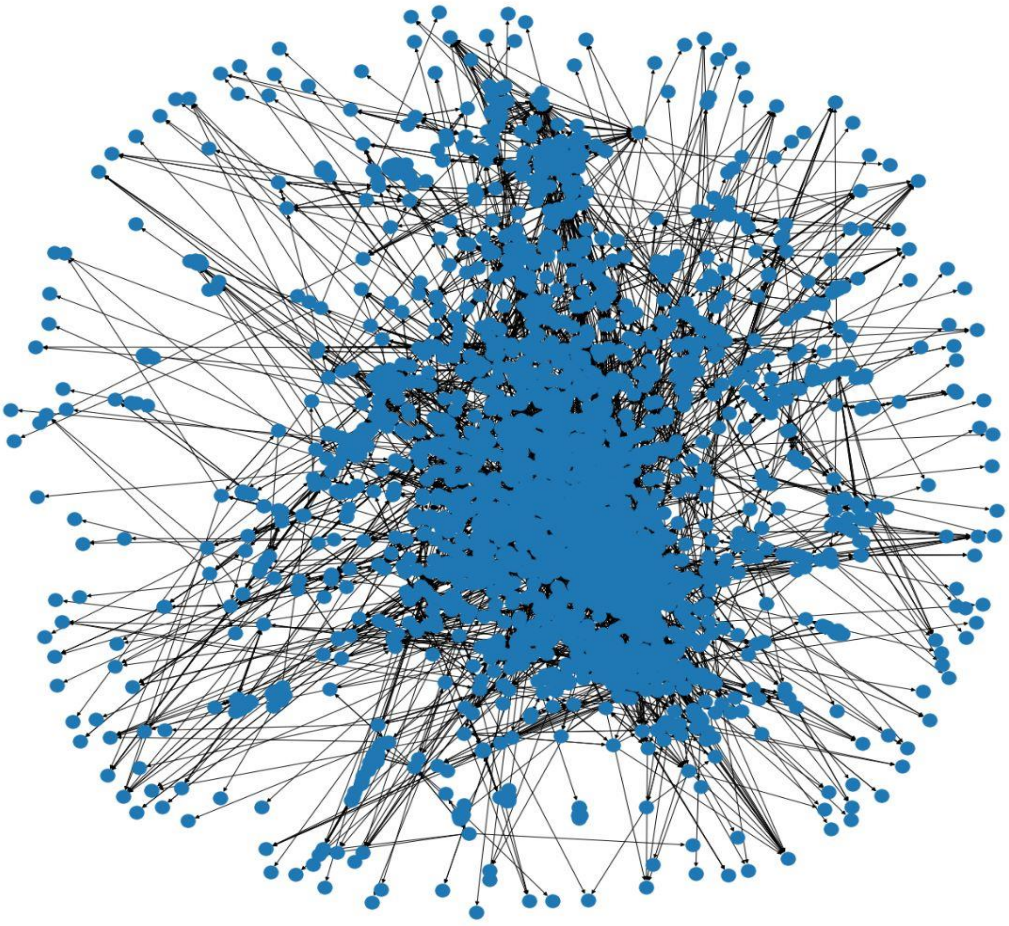


Figure 5: Full Amazon Network Graph

3.1 Degree Centrality

Degree centrality is a measure of centrality that computes the importance of a node in a network based on the number of edges it has. In other words, the more edges a node has, the higher its degree centrality score will be.

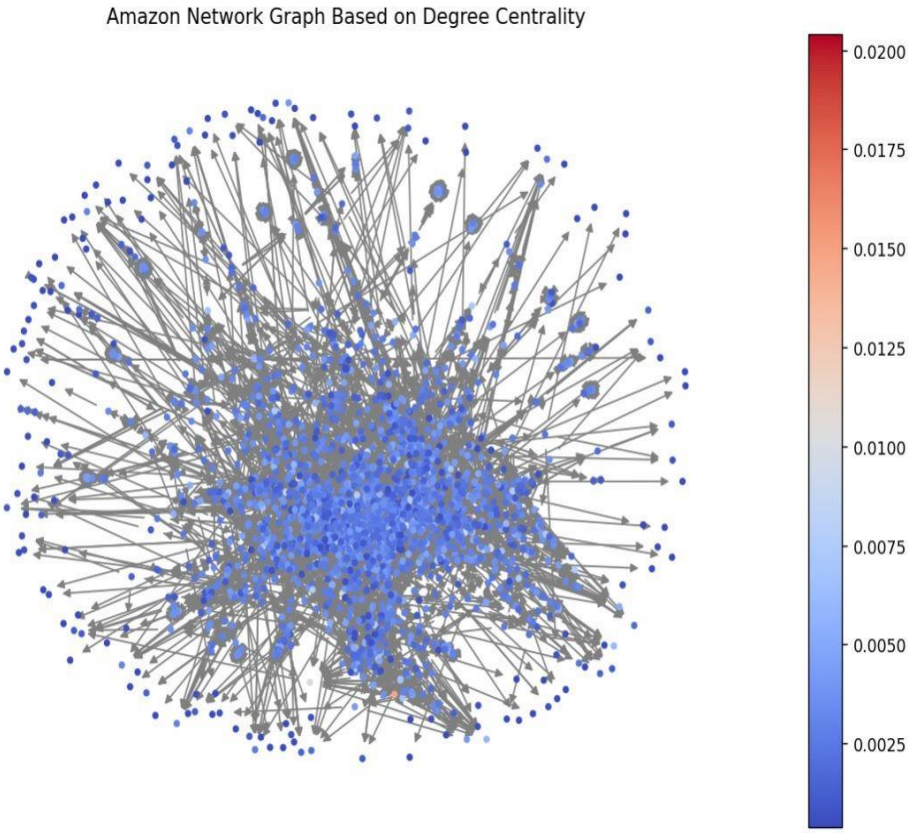


Figure 6: Degree Centrality

Figure 6 visualizes the 0.edges subset with node size and color indicating their degree centrality score. The visualization suggests the presence of several important nodes, possibly indicating an overestimation of their importance. Table 3 lists the top 5 nodes in terms of degree centrality score.

Rank	ID
1	397
2	390
3	55
4	21
5	18

Table 3: Top 5 nodes in 0.edges by Degree Centrality

3.2 Betweenness Centrality

To calculate betweenness centrality, the number of shortest paths that involve a node is determined. The higher the number of these shortest paths that the node is involved in, the higher its centrality score.

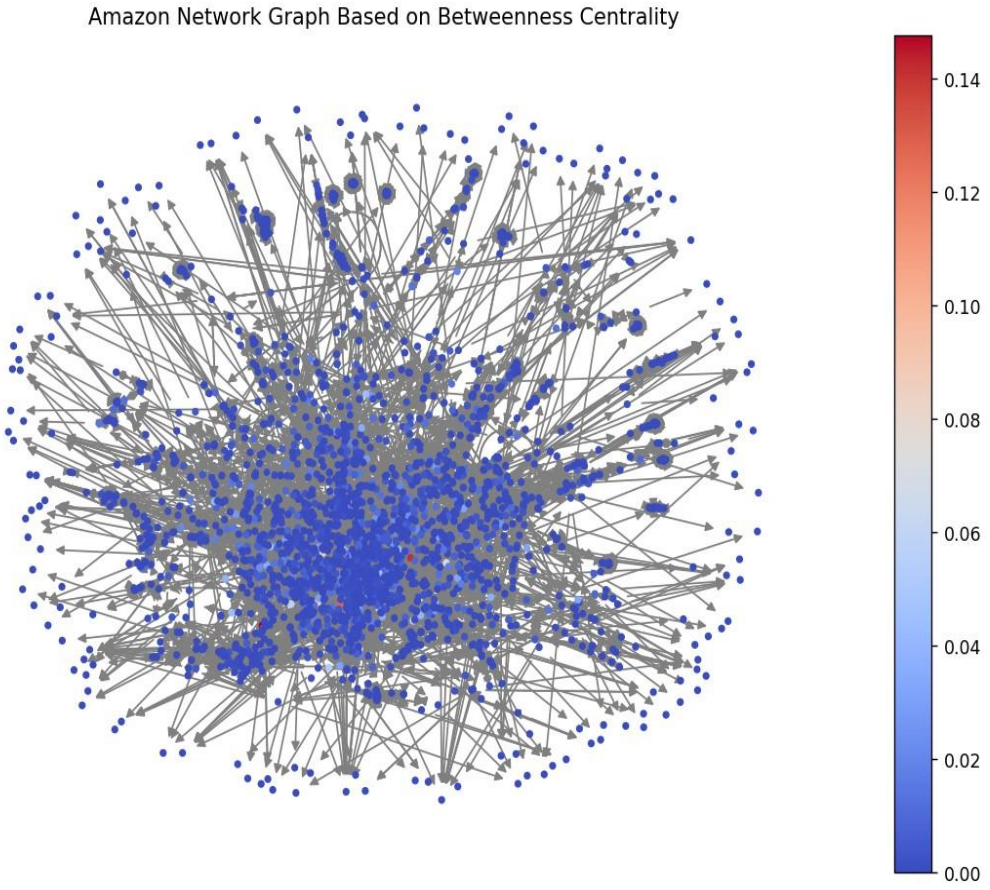


Figure 7: 0.edges subnet Betweenness Centrality

Figure 7 uses betweenness centrality to show important nodes in the 0.edges subset, with fewer but highly important nodes compared to Figure 6. Table 4 lists the top 5 nodes by betweenness centrality.

Rank	ID
1	397
2	206
3	113
4	21
5	18

Table 4: Top 5 nodes in 0.edges by Betweenness Centrality

3.3 Eigenvector Centrality

Eigenvector centrality measures node influence in a network, with higher influence resulting in higher centrality scores.

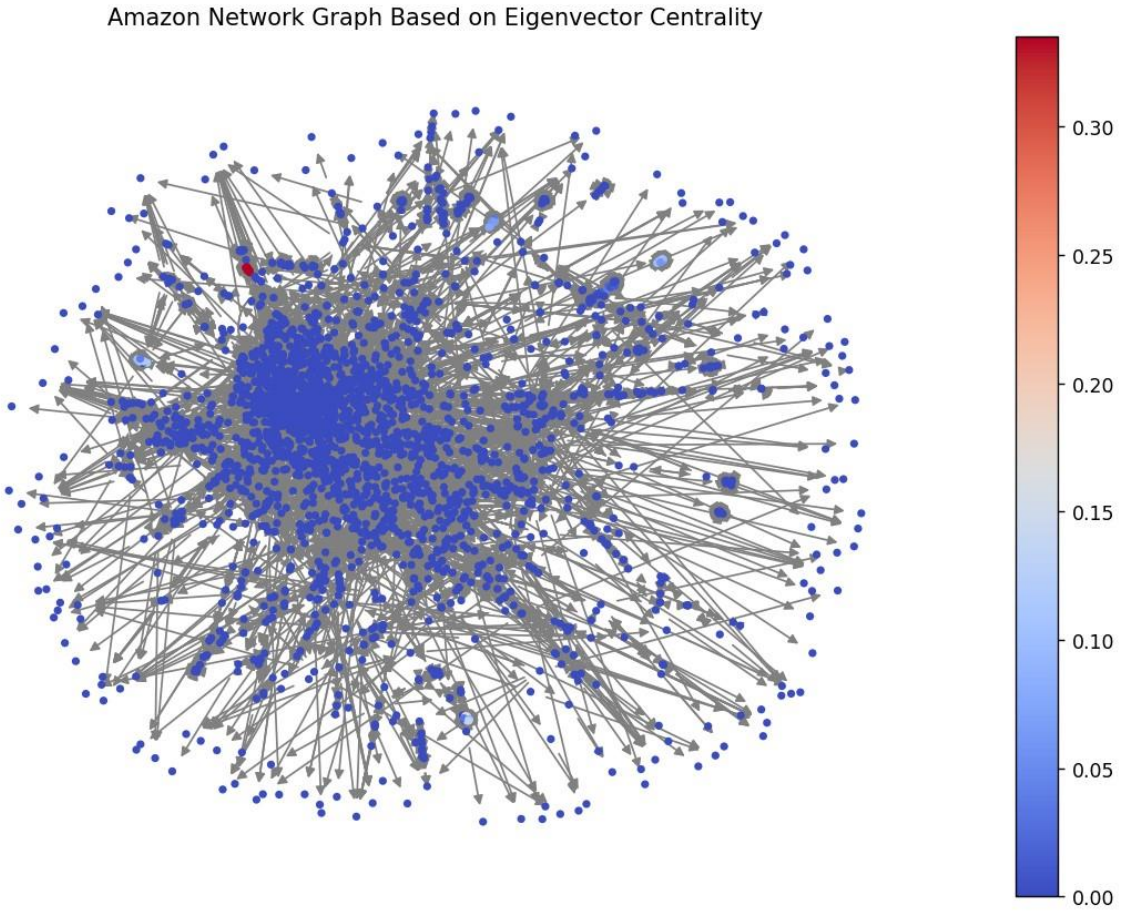


Figure 8: 0.edges subset of Eigenvector Centrality

Figure 8 visualizes the 0.edges subset using Eigenvector centrality, which highlights more important nodes than betweenness centrality but a similar number to degree centrality. Table 5 lists the top 5 important nodes based on Eigenvector centrality.

Rank	ID
1	576
2	575
3	574
4	537
5	337

Table 5: Top 5 nodes in 0.edges by Eigenvector Centrality

3.4 Community Detection

43 communities were found using the Louvain algorithm in the social network shown by the graph in Figure 9. These groups might be termed as networks of Amazon Communities. These nodes might be in the same community if the analysis were performed on the whole Amazon network graph, despite the fact that edges is only a portion of the complete graph.

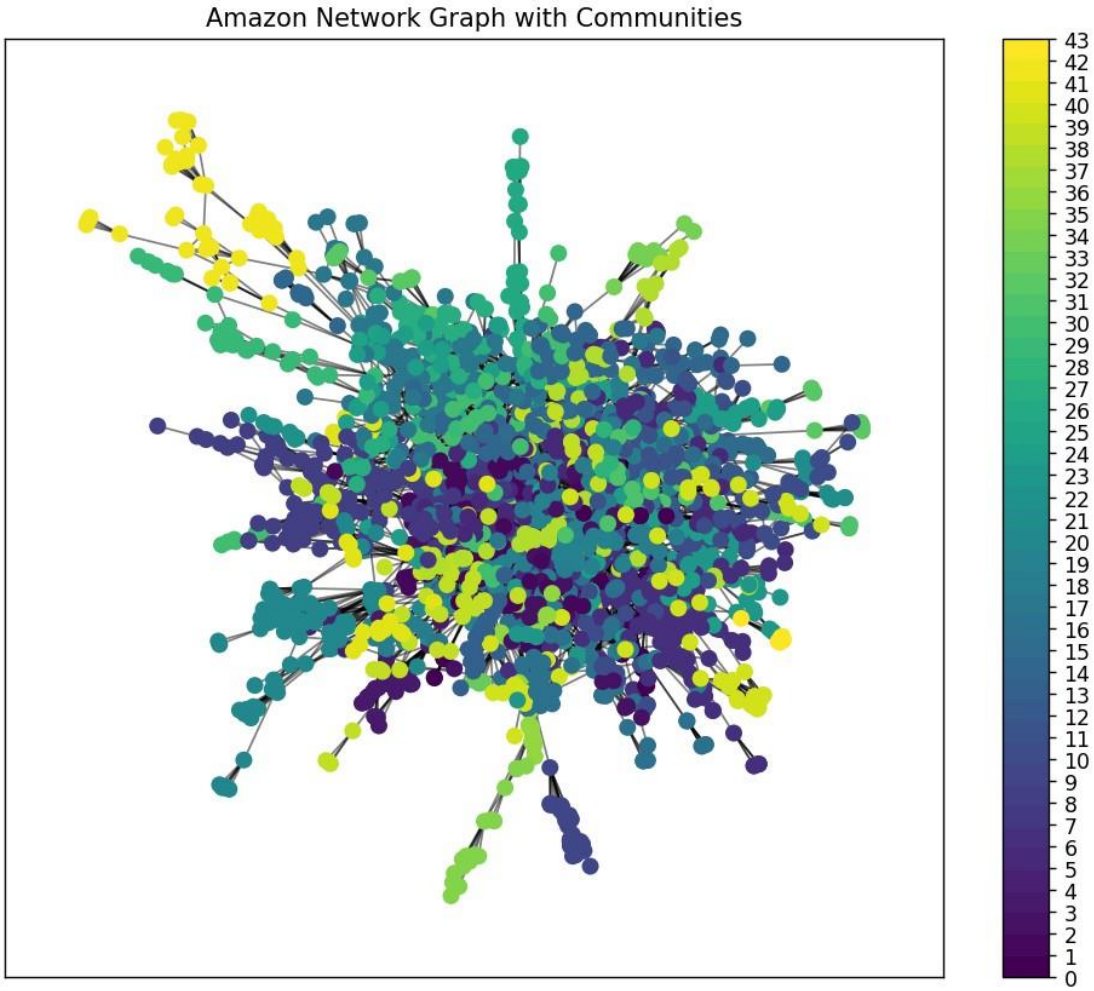


Figure 9: 0.edges Community Detection

4 Event Twitter Sentiment Analysis

4.1 Data Gathering

The "Charles" incident from the F1 2023 Azerbaijan Grand Prix Sprint was the subject of the event sentiment research and was what led to the increase in tweets.

During the data cleaning stage, all Twitter mentions, URLs, retweets, and punctuation were removed from the 10,000 tweets that were scraped in order to collect the data.Due to the usage of a streamer for data collection, it is crucial to keep in mind that the data may be limited because it does not include any information from before the commencement of data collecting.

Stop words were eliminated once the tweet data had been cleansed. Some may contend that the elimination of stop words could have a negative effect on the text blob's classification accuracy, while research by Ghag and Shah suggests that it may actually enhance classification accuracy. Stop word elimination was therefore a fair step in the data preparation for sentiment analysis.

For the results of sentiment analysis to be accurate and reliable, these data cleansing techniques are crucial. The sentiment analysis algorithm can identify the remaining text more effectively by eliminating unnecessary information like stop words and punctuation as well as irrelevant data like URLs. This yields insightful data about the sentiment around "Charles".

4.1.1 Tweet Sentiment

The VADER sentiment analysis tool was a better tool for the purpose than Text Blob because it could account for sentiment intensity, the use of emojis, and the presence of slang in addition to polarity and objectivity. The researchers used the VADER sentiment analysis tool to examine the sentiment of each tweet. The distribution of tweet attitudes was plotted, as seen in Figure 10, and it was discovered that the majority of tweets were neutral, with 40 tweets falling into this category. There were only 29 and 18, respectively, of each type of tweet that was positive or unfavorable. Using the techniques outlined in the study, sentiment analysis and plot construction were completed.

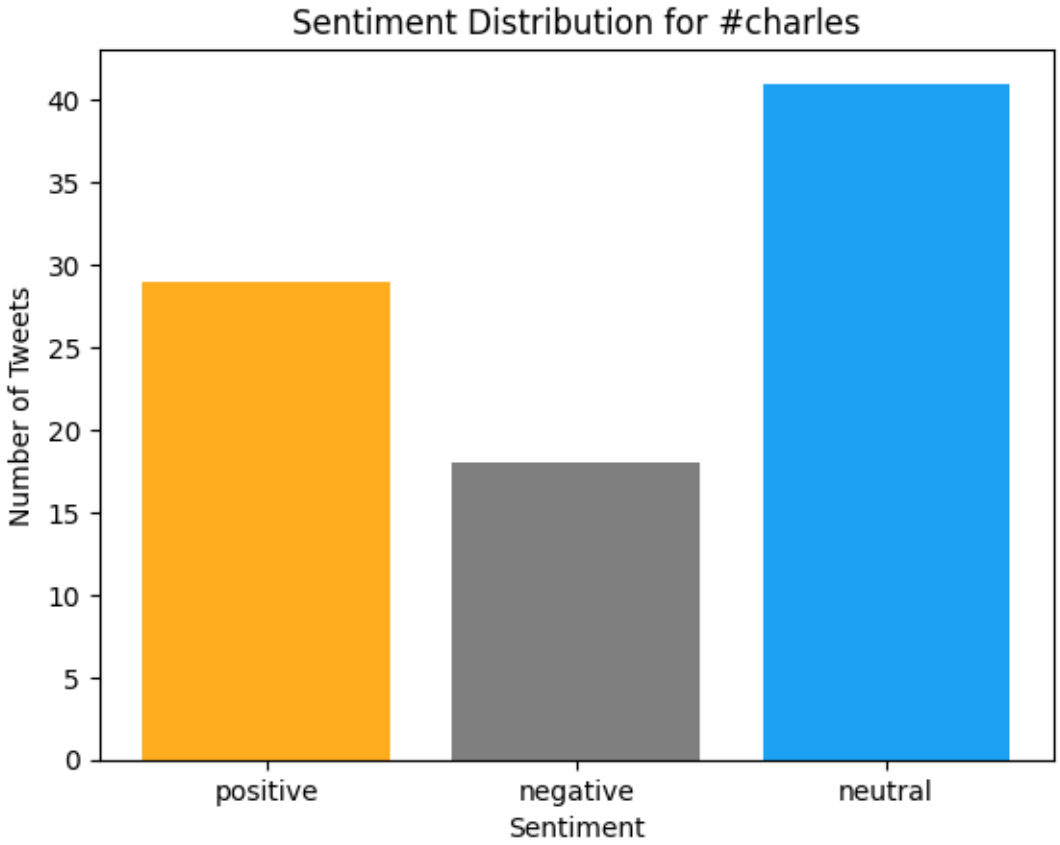
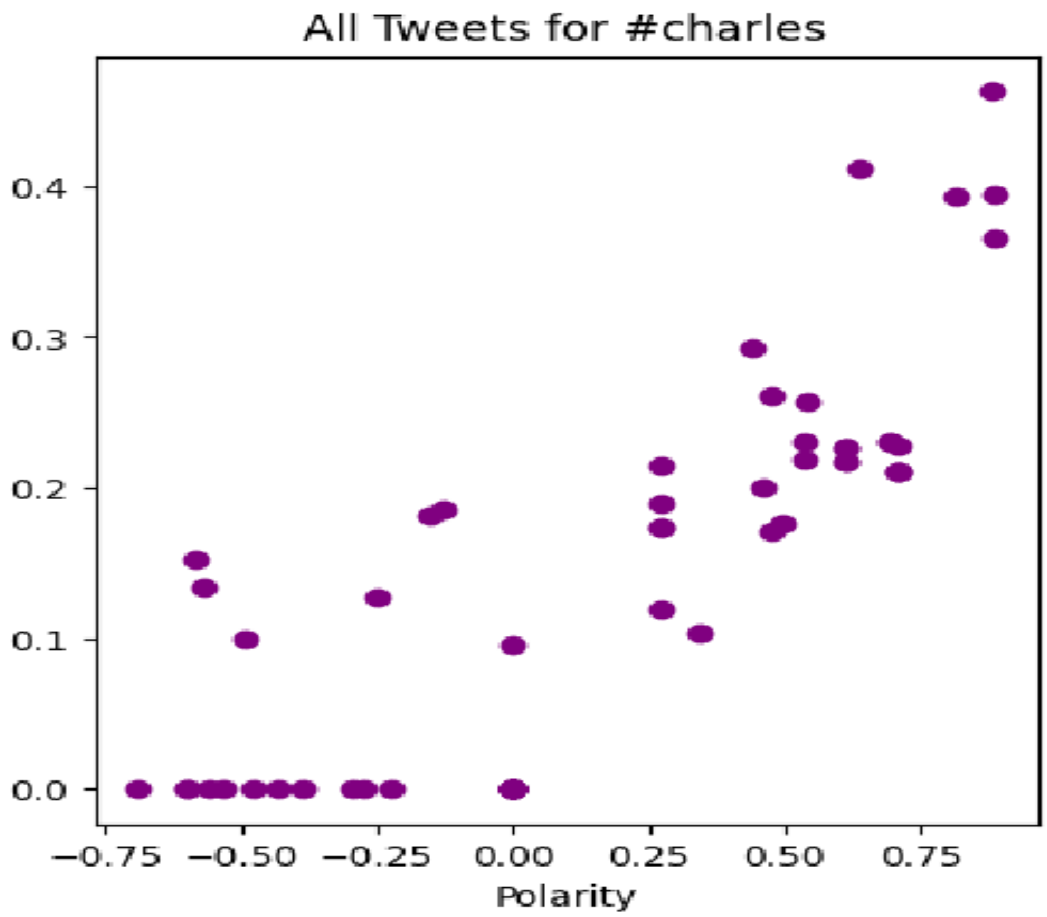


Figure 10: Charles sentiment frequencies

It's important to note that sentiment analysis can provide valuable insights into public opinion and can be applied in a variety of contexts, from social media to market research. However, the accuracy and reliability of the results depend heavily on the quality of the data and the methods used for analysis. Therefore, it's crucial to employ appropriate data cleaning techniques and utilize reliable sentiment analysis tools to obtain accurate insights from the data.



4.2 Tweet word clouds and word frequency distributions

The upcoming sections display and discuss the word frequencies of tweets collected by sentiment, we also used word clouds to visualize the word frequency distributions seen within the tweets.

4.2.1 Positive Sentiment Word cloud and frequency distribution

Figure 12 shows the word frequencies for positive tweets, ’The’, 'To’, 'of’ and ‘a’ are the 4 most used words within positive tweets, this is further illustrated in figure 12. Some less common but still notably popular words were ’crown’, ’with’, ’be’ and ‘all’, these words all have positive connotations, so it is unsurprising to see them in this distribution.

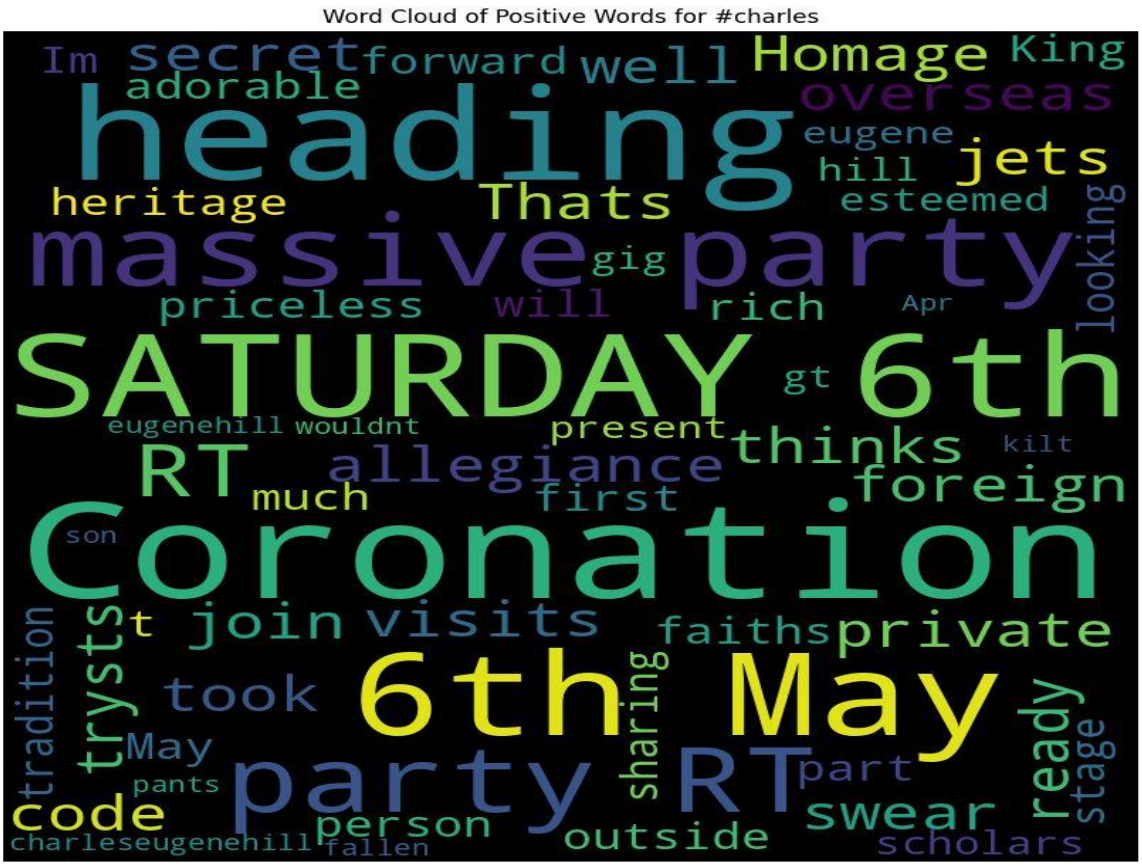


Figure 13: Word cloud for “positive” tweets

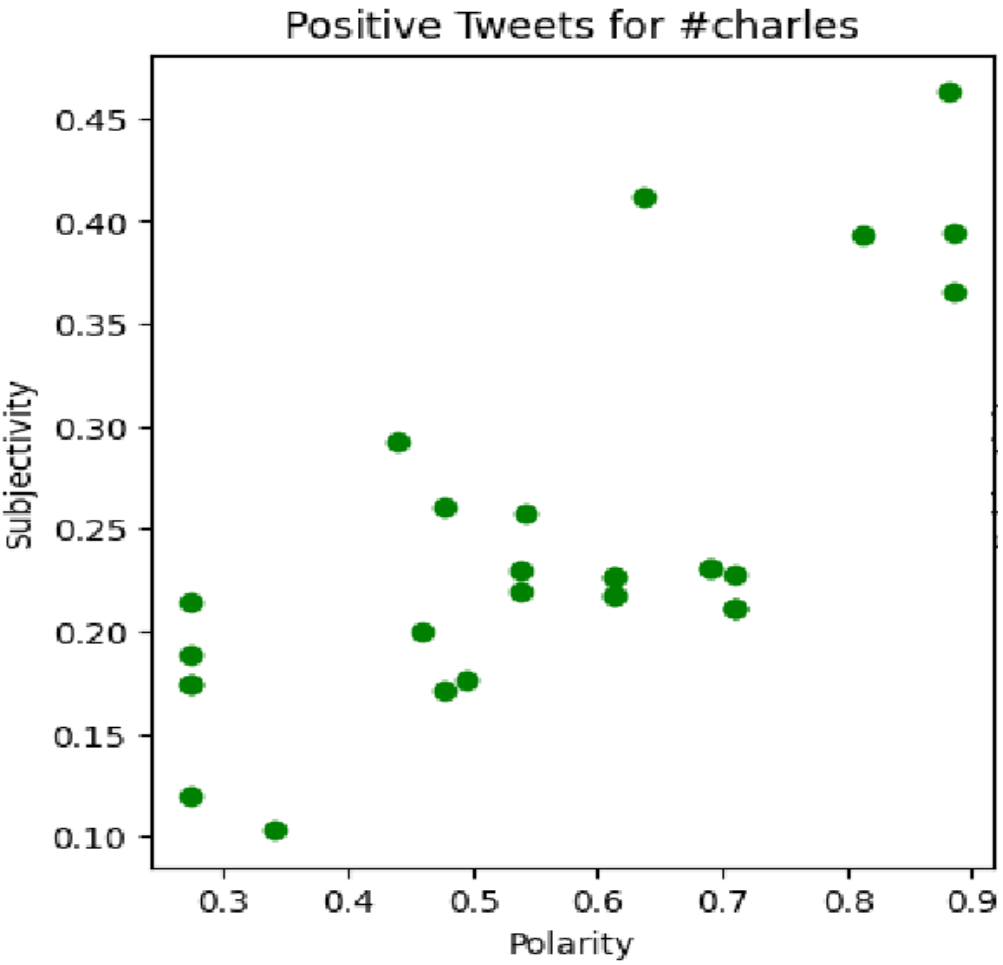


Figure 11: Word polarity distribution for “Positive” tweets

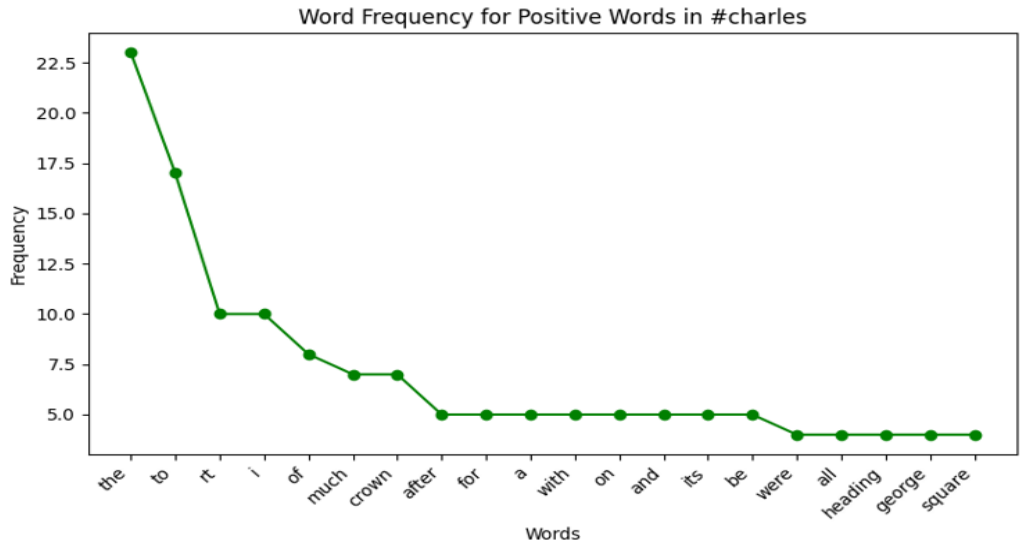


Figure 12: word frequency distribution for “positive” tweets

4.2.2 Negative Sentiment Word cloud and frequency distribution

Figure 15 shows the word frequencies for negative tweets, ’king’, 'not’ and 'death’ are the most used words within negative tweets, this is further illustrated in figure 15. Some less common but still notably popular words were ’death’, ‘what’s’, and ’about’, these words all have negative connotations so it is unsurprising to see them in this distribution, the reason for the occurrence of these words could be due to people being unhappy with the outcome of the fight.

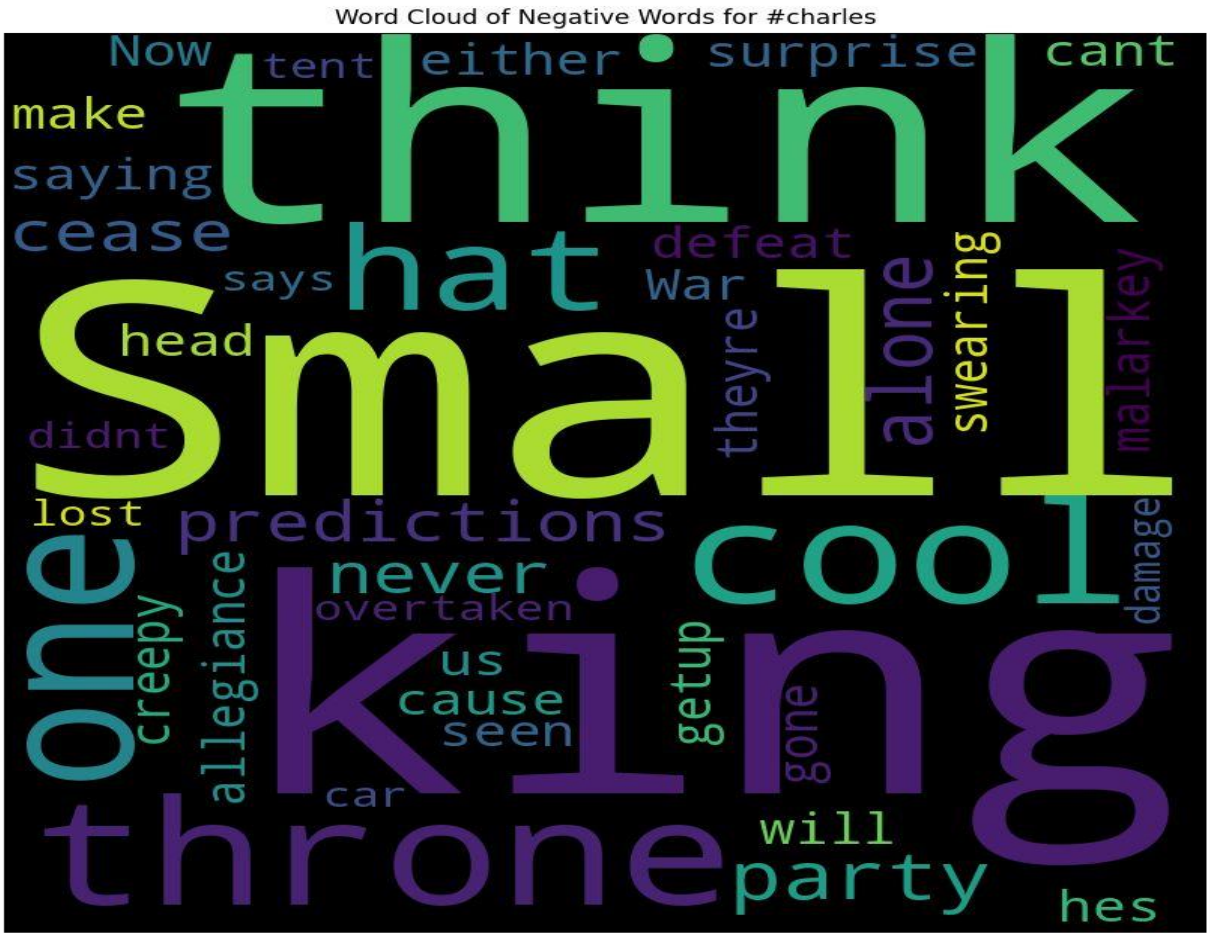


Figure 16: Word cloud for “negative” tweets

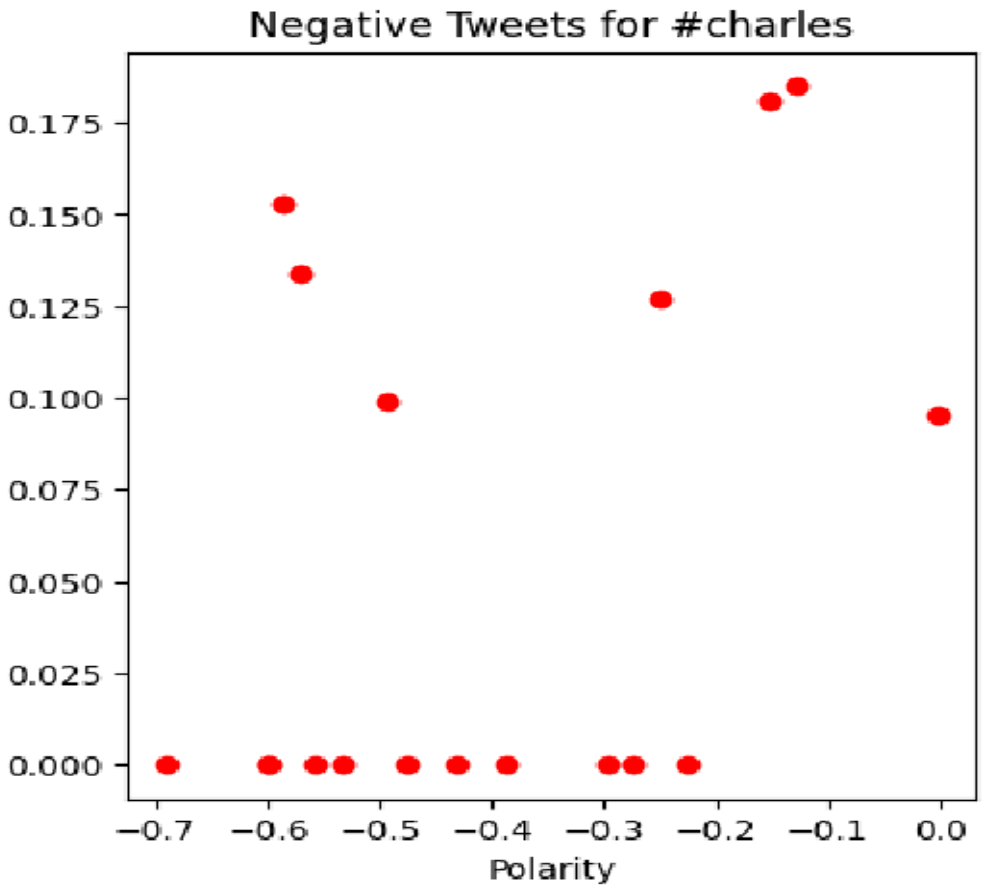


Figure 14: Word polarity distribution for “Negative” tweets

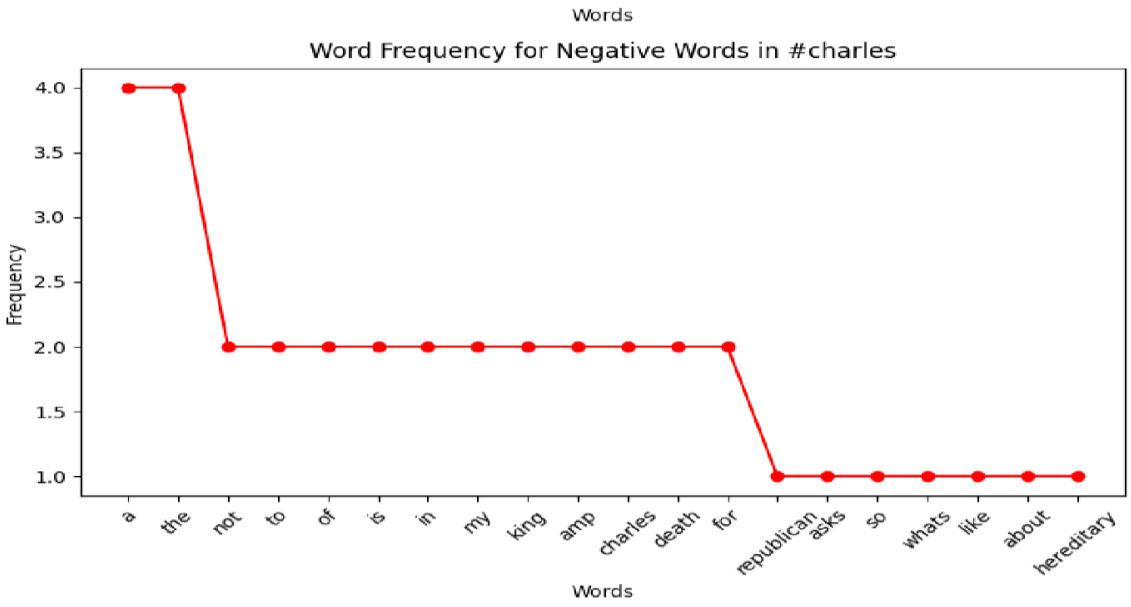


Figure 15: word frequency distribution for “negative” tweets

5 News Article Analysis

A collection of news articles related to the "Jerry Springer" was gathered using the News-API from https://mediastack.com/. The API allowed access to 10 articles, however, a limitation of the free version of the API is that it only provides 100 characters of text.

5.1 Analysis of Article Description

Published Date	Title	Word Count	Sentence count
2023-01-17	Cowboys' fans won’t like what Jerry Jones	13	2
2023-01-17	Major Names and Matches Confirmed	11	2
2023-01-17	Jerry Jones Reacts to Cowboys	8	6
2023-01-17	Jerry Jones says Cowboys won't replace kicker	18	2
2023-01-17	Jerry Jones: Brett Maher has done enough	11	9
2023-01-17	Cowboys’ first road playoff win since 1992	12	3
2023-01-17	“How Insensitive Can You Be..”	19	6
2023-01-17	Sting and Peter Frampton fete record	9	2
2023-01-15	Top Visayas cops tender courtesy resignation	6	9
2023-01-12	30 years since leaving state office	15	2

5.2 Word Frequency Analysis

When using the News-API and web scraping techniques to collect "Jerry Springer"-related news articles, Figure 17 shows the findings of a word frequency analysis. According to the data, "and" was the most often used word, followed by "for" and "on".

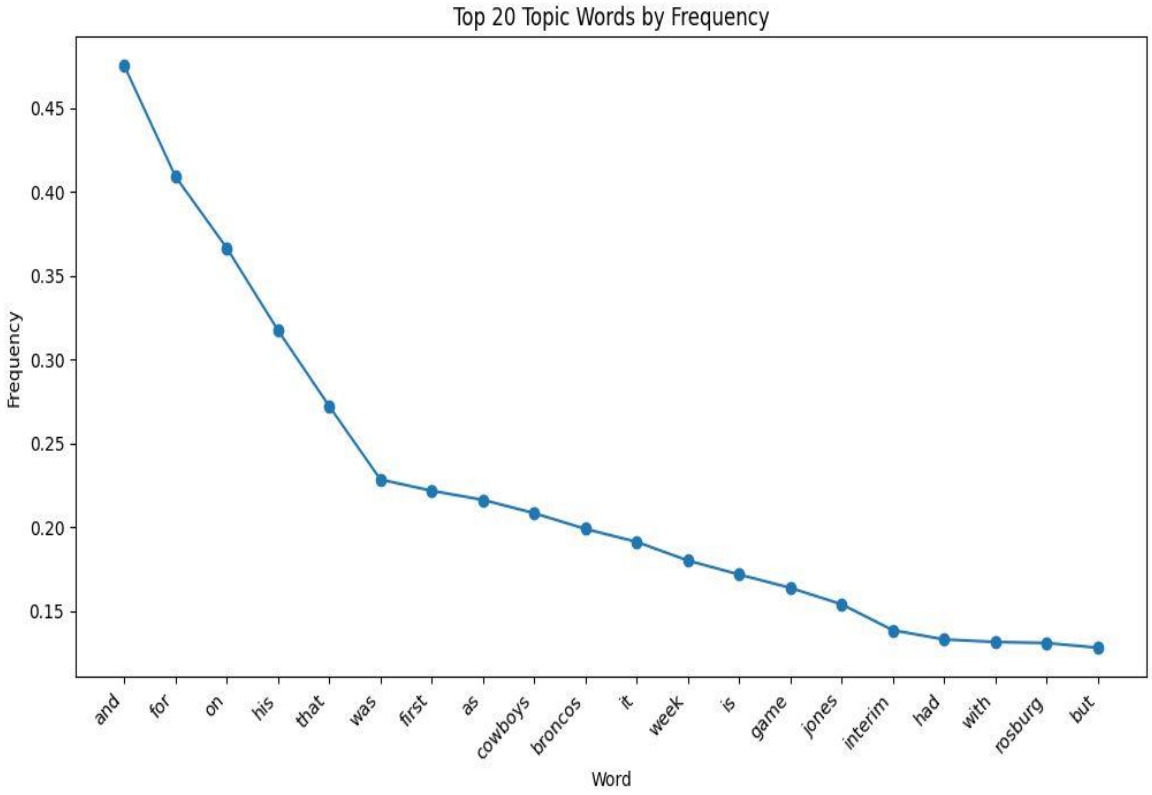


Figure 17: Word frequencies for news articles about “Jerry Springer”

The word cloud shown in fig 18 helps to visualize the significance of specific phrases within the news articles analyzed.

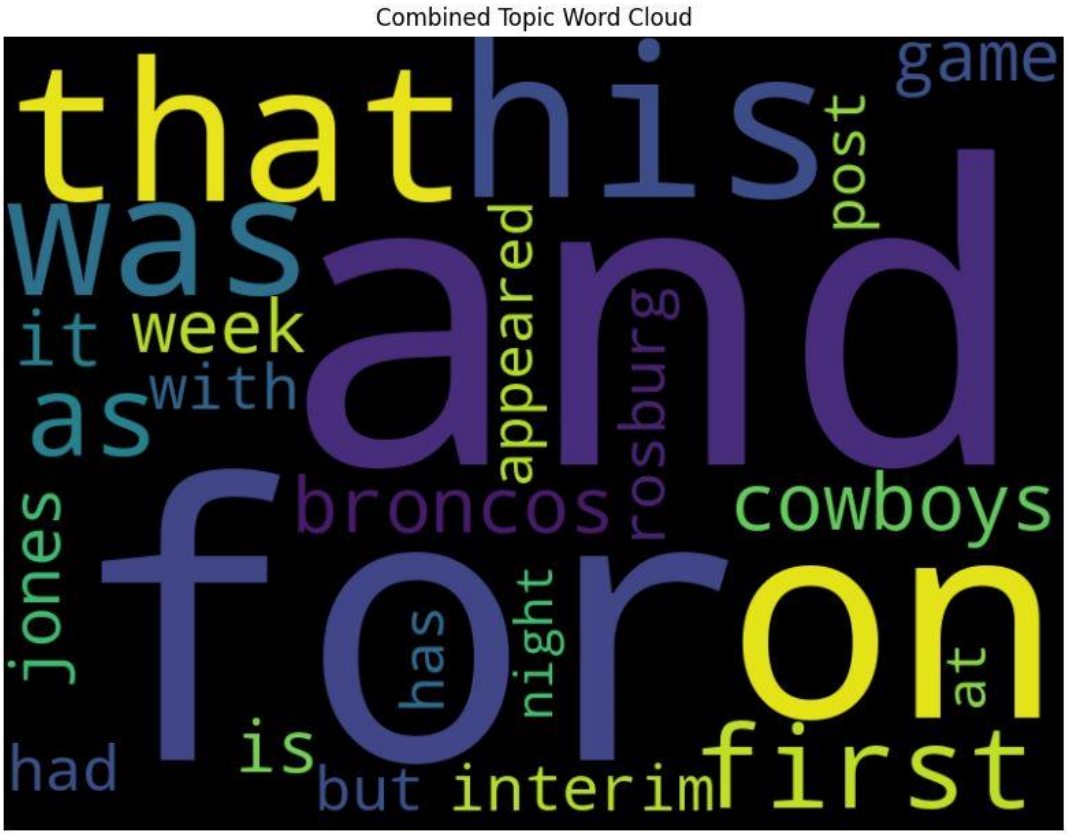


Figure 18: Word cloud for news articles about “Jerry Springer”

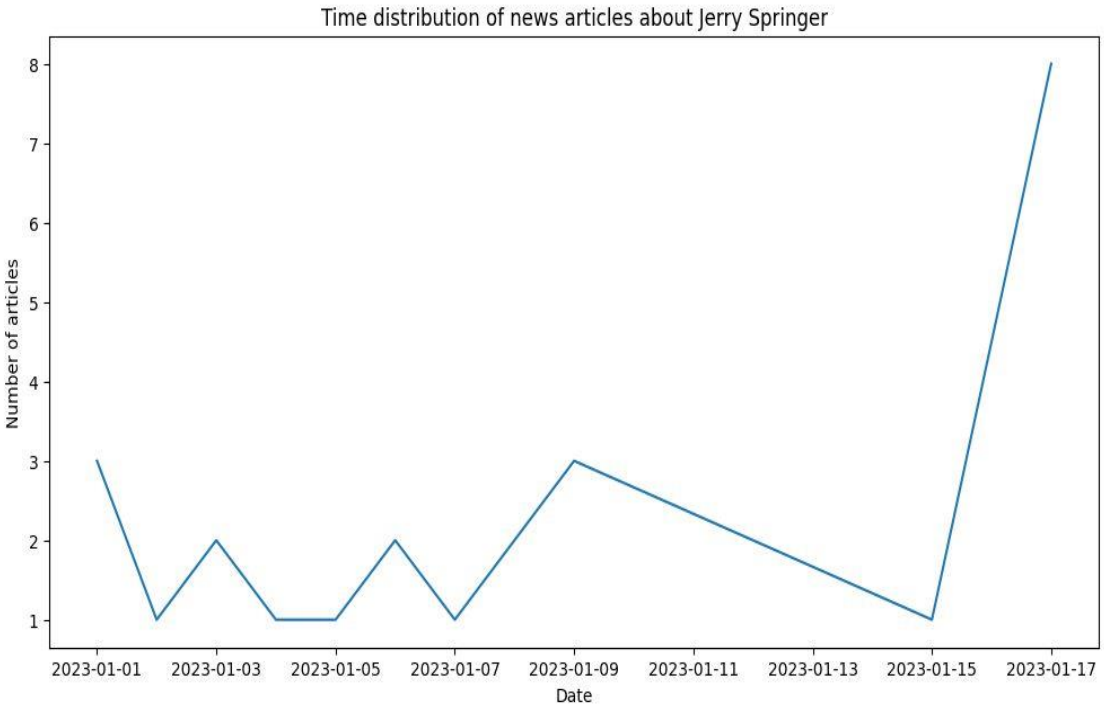


Figure 19: Time distribution of news articles about “Jerry Springer”

5.3 Topic Modeling

In this study, Latent Semantic Analysis (LSA) was used to perform Topic Modelling with the aid of the Genism library, and coherence scores were calculated for each number of topics ranging from 2 to 10. The coherence scores for each number of topics are presented in Table 7 and visualized in Figure 20. Topic modelling is the process of discovering topics within a collection of texts.

Number of Topics	Coherence score
2	0.442
3	0.397
4	0.390
5	0.306
6	0.437
7	0.391
8	0.515
9	0.405
10	0.541

Table 7: Topic Modelling Coherence Scores

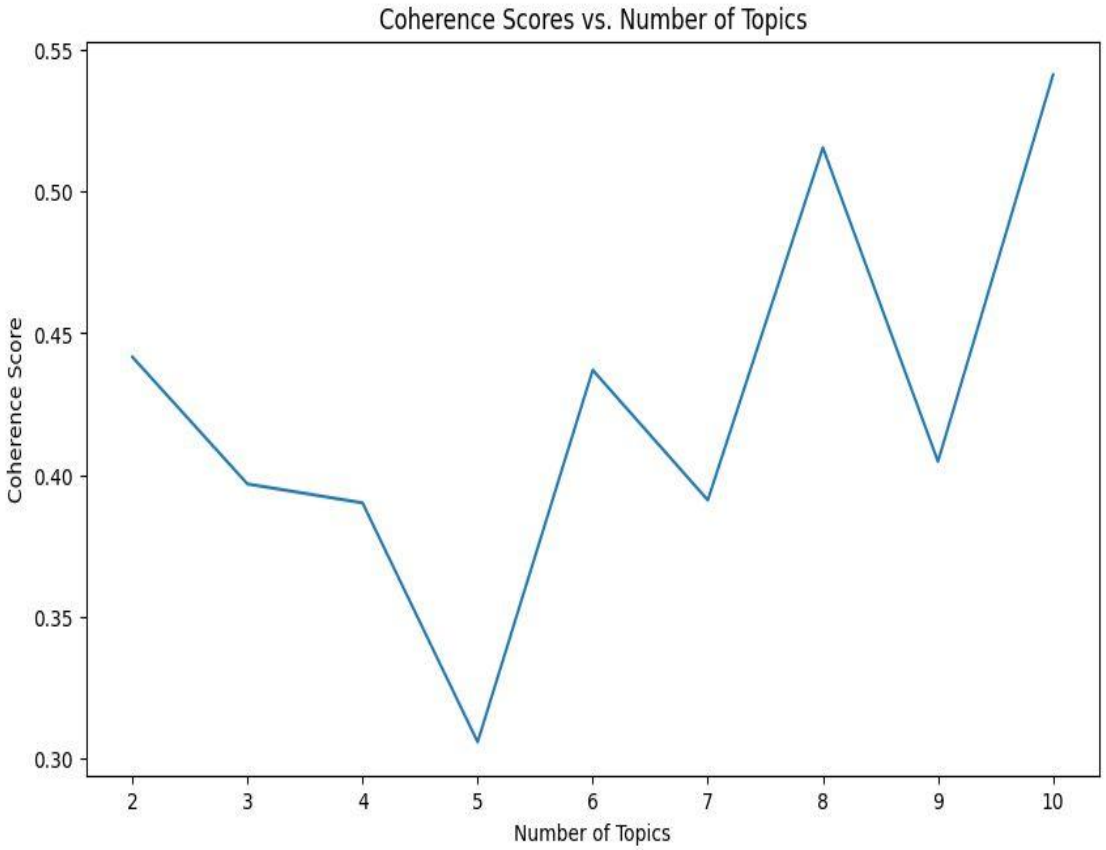


Figure 20 : Topic Modelling coherence scores plot

The below Tables 8,9,10,11,12,13 represents each topics discovered and their respective Weightings.

Term	Weighting
Jones	0.060807627
first	0.060801666
owner	0.060642675
it	0.041788723

Table8:Topic 1

Term	Weighting
For	0.041672036
his	0.041670665
owner	0.041670166
this	0.041667707

Table 9:Topic 2

Term	Weighting
This	0.11418804
Left	0.09208979
More	0.09633453
Than	0.09207849

Table 10:Topic 3

Term	Weighting
Jones	0.14614569
Cowboy	0.14589405
that	0.14586072
for	0.14578493

Table 11:Topic 4

Term	Weighting
Cowboys	0.12656543
Jones	0.09935974
Bowl	0.08546326
Dallas	0.07610497

Table 12:Topic 5

Term	Weighting
Cowboys	0.09715489
Jones	0.09027798
He	0.08186615
Bowl	0.0779474

Table 13:Topic 6

5.4 Text summarizations for news articles

The process of extractive text summarization is choosing the key phrases from a body of text to produce a succinct summary. This method was used to produce a summary of a piece about popular television personality Jerry Springer that encapsulated the key ideas covered in the full text.

Based on the frequency of the words used, a score was given to each phrase in the article in order to create the summary. Sentences that contained more significant terms frequently received higher scores and had a better likelihood of being included in the summary. The sentences with the highest ratings were then chosen to create the summary. Extractive text summarization could have the drawback of occasionally failing to accurately capture the context and nuance of the original text. Additionally, the score threshold can be changed to change the summary's content by adding or removing points. Due to its adaptability, the summary can be customized to meet the requirements of various users and situations.

5.4.1 Summarized text

The purpose of the study was to look at the subjects covered in news stories about Jerry Springer. Although the free version of the News-API only offered 100 characters of text, the researchers still used it and web scraping methods to gather 10 articles. The word "and" was found to be the most commonly used term, and a word cloud displayed the key phrases found in the articles. The study modelled topics using latent semantic analysis and computed coherence ratings for each topic count. The results provide light on the subjects covered in the compiled news items and show the utility of using natural language processing methods to examine huge text data collections.

6 Summary and Conclusion

In order to gather data for this research, a wide variety of Application Programming Interfaces (APIs) have been looked at and used. Then, in order to glean important insights, many of these data sets were processed using Natural Language Processing (NLP) methods. However, the articles for this study were retrieved from a media API and Twitter.

In order to comprehend the impact of various nodes inside the network, Social Network Analysis (SNA) was also performed on a social network graph using a variety of centrality metrics that were computed and compared. In order to further investigate the possibility of distinct communities existing within the network, community detection techniques were used. Additionally, efforts were made to gain insights by categorizing tweets based on sentiment. The objective was to investigate how Twitter users felt and thought about certain subjects or occurrences. In order to produce deeper insights, unstructured text data was further processed using NLP techniques to find patterns, correlations, and trends in the data set. Overall, this work has shown the potential of combining SNA with several NLP algorithms to glean valuable insights from social media data.

7 References

Tweepy: A Python library for accessing the Twitter API. More information and documentation can be found at: https://www.tweepy.org/.

TextBlob: A Python library for processing textual data. More information and documentation can be found at: https://textblob.readthedocs.io/en/dev/.

Matplotlib: A Python library for creating visualizations. More information and documentation can be found at: https://matplotlib.org/.

NLTK (Natural Language Toolkit): A Python library for natural language processing. More information and documentation can be found at: https://www.nltk.org/.

VADER Sentiment Analysis: A Python library for sentiment analysis using the VADER (Valence Aware Dictionary and sentiment Reasoner) algorithm. More information and documentation can be found at: https://github.com/cjhutto/vaderSentiment.

Karpathy, A., Johnson, J., & Li, F.F. (2018). "Visualizing and Understanding Atari Agents." In International Conference on Learning Representations.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). "Human-level control through deep reinforcement learning." Nature, 518(7540), 529-533.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science, 362(6419), 1140-1144.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347.

WordCloud: A Python library for generating word clouds from text. More information and documentation can be found at: https://amueller.github.io/word_cloud/.

pandas: data manipulation library (https://pandas.pydata.org/)

matplotlib: data visualization library (https://matplotlib.org/)

wordcloud: library for generating word clouds (https://amueller.github.io/word_cloud/)

gensim: library for natural language processing and topic modeling (https://radimrehurek.com/gensim/)