Web Social Media Analytics and Visualization

Final Submission

Student ID: 23206188

Student Name: Dinesh Thapa

Part of a

Web Social Media Analytics - CMP7203

BIRMINGHAM CITY UNIVERSITY

FACULTY OF COMPUTING ENGINEERING AND THE BUILT ENVIRONMENT

BIRMINGHAM CITY University

Table of Contents

Table of Figures

List of Tables

Statistical Analysis

Statistical Analysis is the collection, interpretation and investigation of data to find trends and patterns to develop valuable insights that helps to make more informed data-driven decisions (Tang and Yang, 2012). Extracting data from social media is now a crucial area because it helps identify key trends for businesses using scientific methods and expertise. The main types of analysis include examining social networks, determining sentiment, and suggesting collaborations (Sapountzi and Psannis, 2018).

1. Twitter Trends Analysis

It is the study of popular topics on Twitter to understand what people are talking about the most to track shifts in public sentiment and opinion over time(Annamoradnejad and Habibi, 2019) This analytical approach is very helpful for understanding how discussions evolve around current events, brands or public figures.

For twitter trend analysis, Covid 19 tweets data is used.

1.1 What are popular trends on Twitter at the moment in the USA?

To discover what's popular on Twitter in the USA, we have to look at the most common hashtags used in the tweets by the USA twitter users. Hashtags often highlight trending topics that gives us insight into what people are discussing and writing on twitter right now (Doshi *et al.*, 2017). This analysis helps us understand the latest public conversations and interest on Twitter.

Table 1: Top 10 Hashtags on Twitter in the USA

Hashtag	**Mentions**
#covid19	13907
#coronavirus	1327
#covid19.	848
#covid19,	551
#covid19…	539
#pandemic	276
#covid19?	231
#trump	224
#wearamask	157
#covid	157

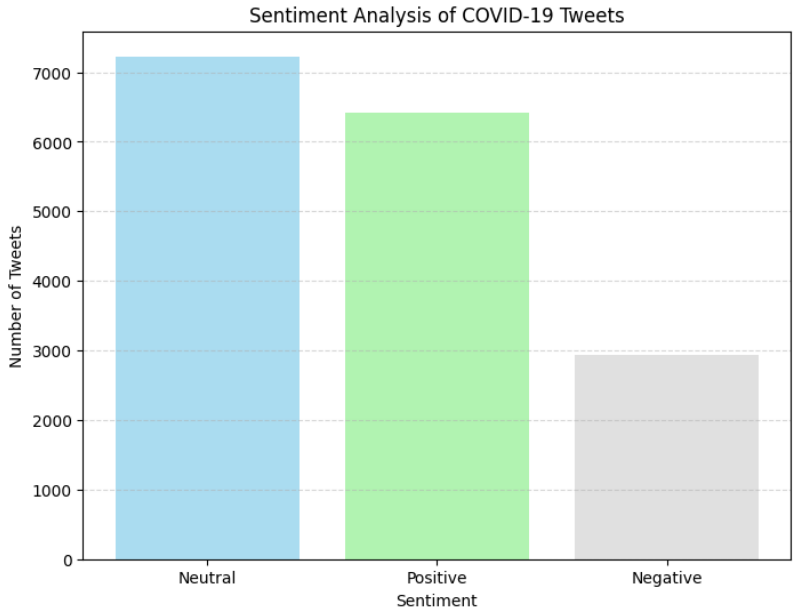
Table 1 shows that #covid19 is the dominant topic with 13,907 mentions that describes extensive discussions about the pandemic. Other related hashtags like #coronavirus, #covid19, and #pandemic also appear prominently which indicates ongoing concern about health crisis. In addition, #trump and #wearamask denotes that people are also discussing about political as well as health safety issues. This gives insights into the current trends and concerns among Twitter users in the USA.

1.2 Extract some insights from these trends such as: when it started in each place?

The first tweet with #covid19 in the US came from Seattle, WA on 2020-07-24 23:47:12.

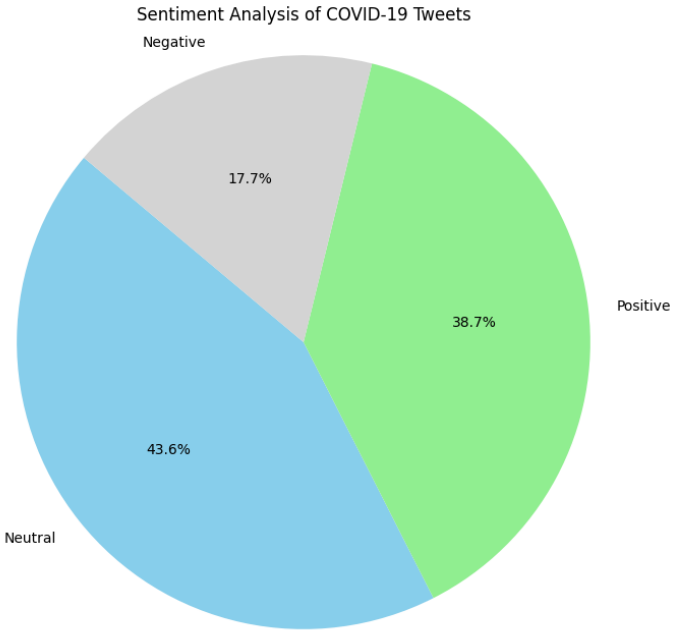
Now, we will perform the sentiment analysis for the hashtag #covid19 to understand the sentiment distribution like Positive, Negative or Neutral from it. In addition, we will also explore multiple locations from which hashtag #covid19 is tweeted in the USA using Heatmap visualization.

Figure 1: Sentiment Distribution of US Tweets containing #covid19



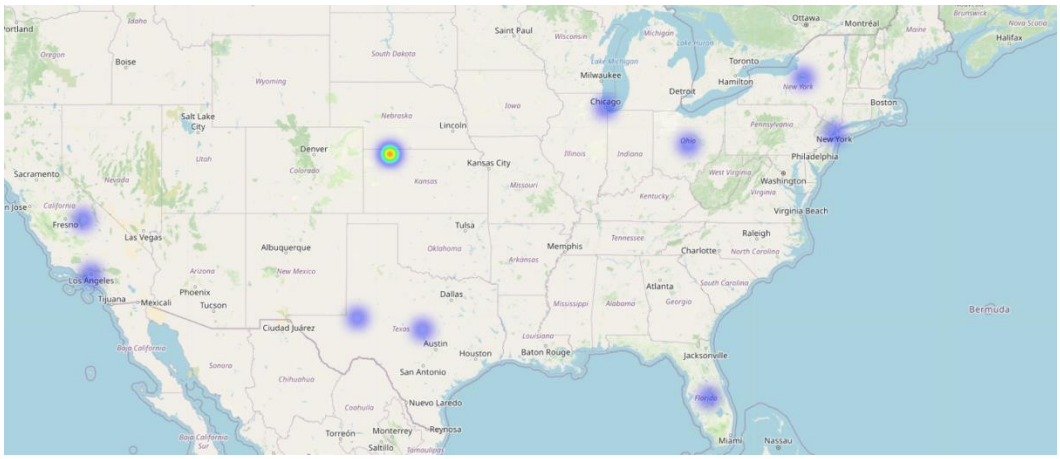
The above bar graph provides the sentiment analysis of tweets that include hashtag #covid19 categorizing the sentiments expressed into three groups: Neutral, Positive, and Negative. The majority of tweets display positive or neutral sentiments that indicates a tendency towards supportive and constructive discussions about Covid-19 on Twitter.

Figure 2: Sentiment Distribution of US Tweets containing #covid19



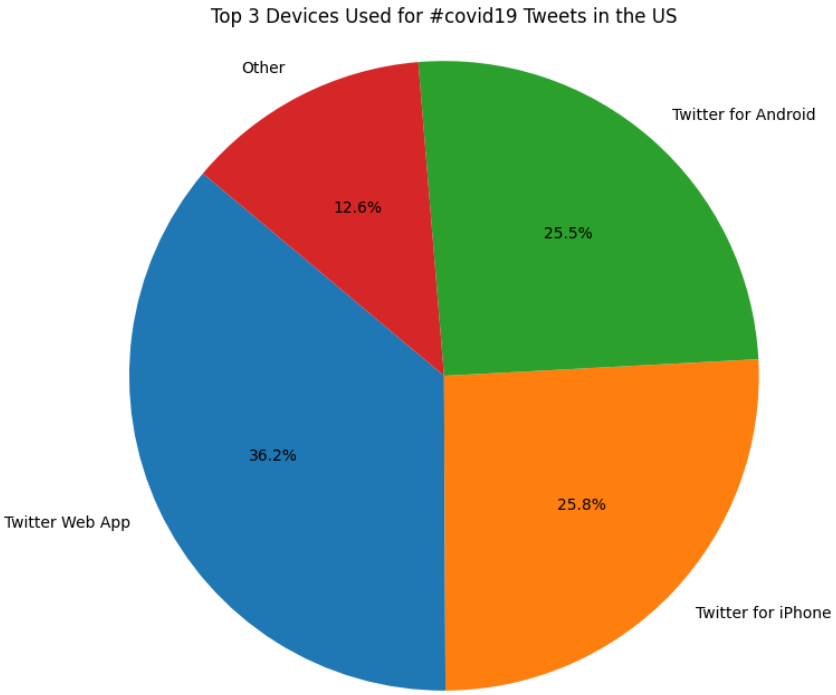
The pie chart shows that Positive tweets constitute the segment at 38.7%, neutral tweets at 43.6%, and Negative tweets at 17.7%.

Figure 3: Locations in the USA from which tweets having #covid19 is posted



This heat map illustrates the distribution of tweets tagged with #covid19 across the USA. Major activities are visible in populous urban regions such as New York, Los Angeles, and Chicago. The presence of tweets across diverse regions indicates nationwide engagement with the topic of Covid-19.

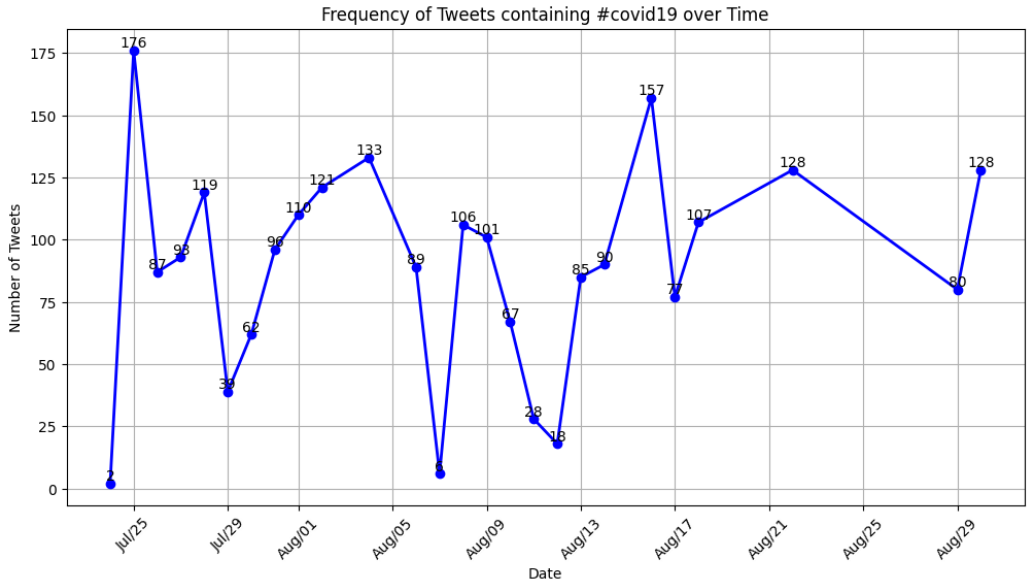
1.3 What devices are used to tweet?



The pie chart illustrates the distribution of devices used for tweeting with the hashtag #covid19 in the United States. Large segment of 36.2% people used Twitter Web App, 25.8% used iPhone, and 25.5% used Android for tweeting. The remaining 12.6% labeled as “Other” could include less common devices such as tablets, third-party apps, or other mobile devices.

1.4 Trends Analysis: Volume of Tweets Over Time in the USA

Figure 4: Volume of Tweets with Hashtag #Covid19 Over Time

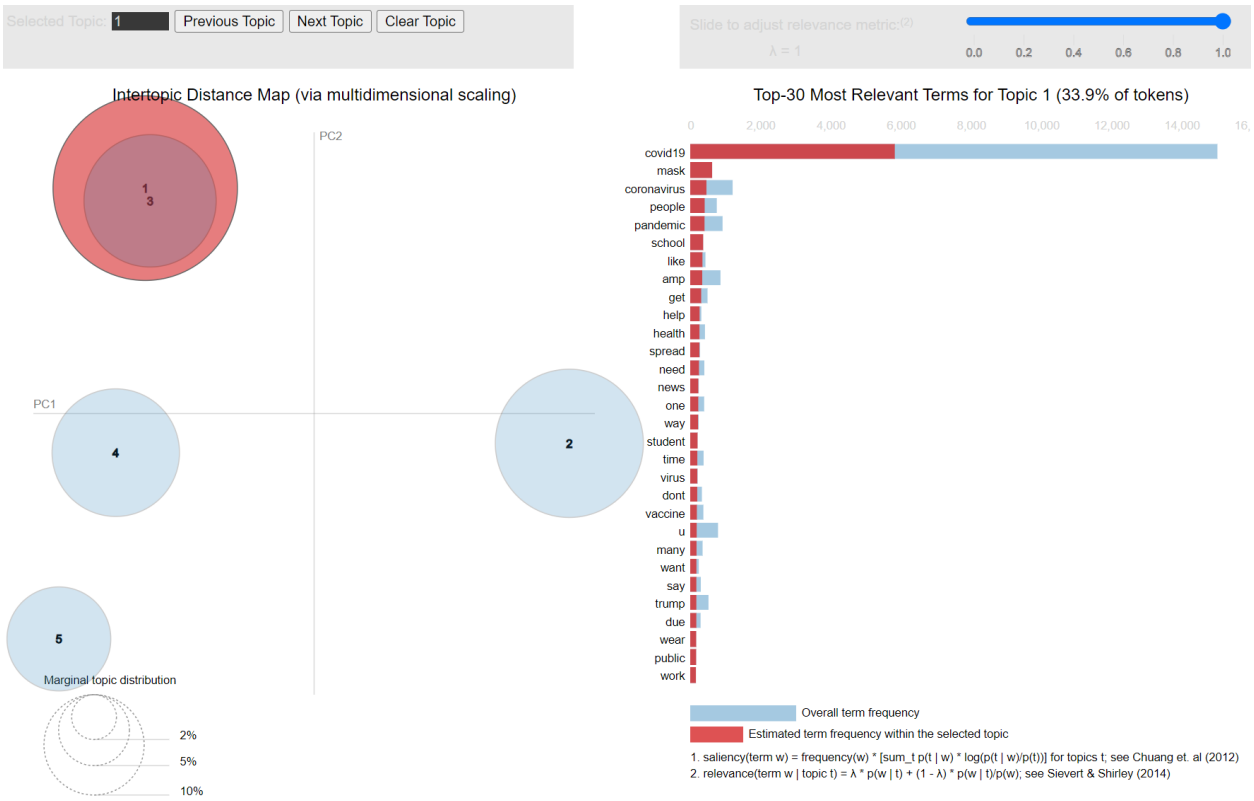


The graph displays the frequency of tweets containing the hashtag #covid19 over a specific period in July and August 2020. It shows significant fluctuations in the number of tweets ranging from as few as 2 to as many as 176 on different days. Peaks in tweeting activity on certain dates suggest heightened interest or events related to COVID-19 while the valleys indicate lesser activity on other days.

1.5 LDA Topic Modeling (Covid19 in USA)

LDA (Latent Dirichlet Allocation) is a statistical model used primarily in natural language processing and machine learning to categorize or group documents into topics. Here, we are going to use the LDA topic modeling to find the topics that are relevant to our dominant words in which each topic is characterized by a distribution of words. LDA helps in discovering the hidden thematic structure in large archives of documents.

Figure 5: LDA Topic Modeling to show Relevant Terms for Topic



Let's visualize the above LDA model topics through Word Cloud visualization for each topic.

Figure 6: Word Cloud Topic 0

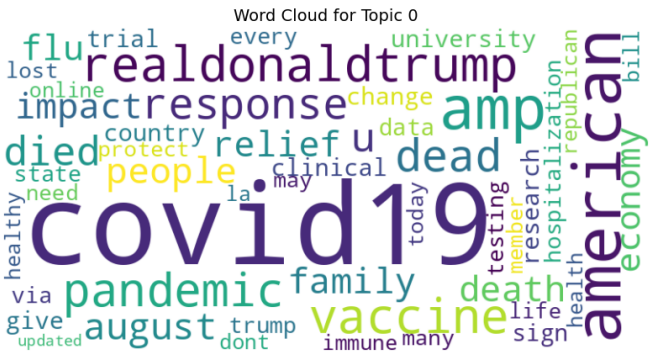


Figure 7: Word Cloud for Topic 1

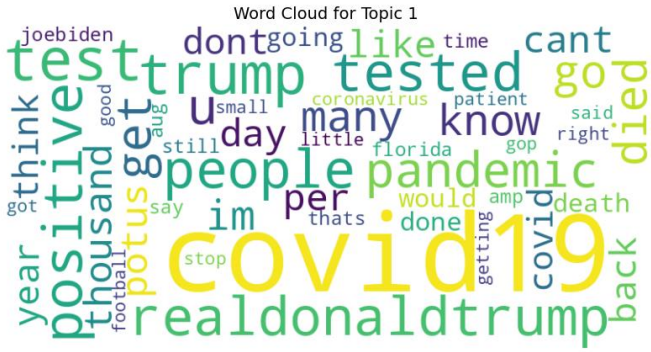


Figure 9: Word Cloud for Topic 2

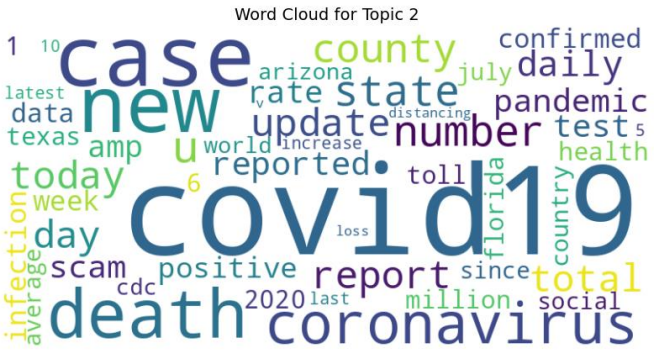


Figure 8: Word Cloud for Topic 3

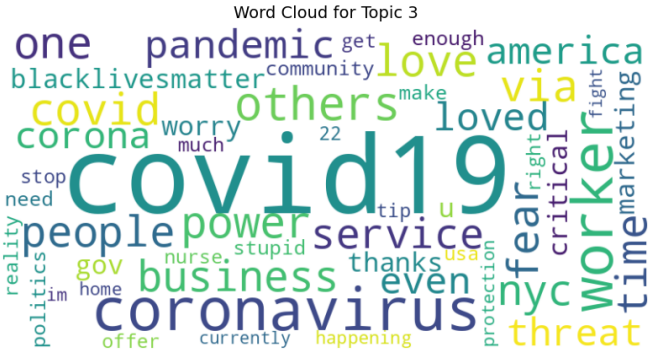
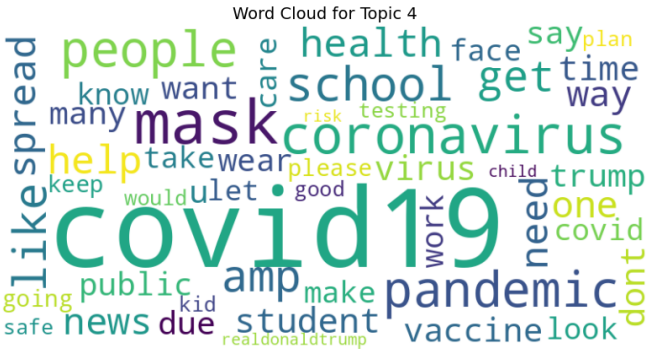


Figure 10: Word Cloud for Topic 4



The provided word clouds represent different topics derived from discussions about COVID-19 in the USA using LDA topic modeling.

Topic 0 focuses on government responses and impacts, highlighting terms like “vaccine”, “pandemic” and political figures such as “realdonaldtrump”.

Topic 1 describes about testing and public sentiment with key terms like “test”, “trump” and “people” indicating discussions around testing accessibility and political influence.

Topic 2 centers on case numbers and statistical reporting with frequent terms including “case”, “new”, “death” and “report” reflecting a focus on the reporting of COVID-19 statistics.

Topic 3 explores the broader social impact and community responses featuring words like “people”, “love”, “business” and “blacklivesmatter” suggesting discussions on how COVID-19 intersects with social movements.

Finally, Topic 4 emphasizes preventive measures and public health with dominant terms such as “mask”, “health”, “school” and “safe” pointing to discussions around safety protocols and the impacts on educational and other public spaces.

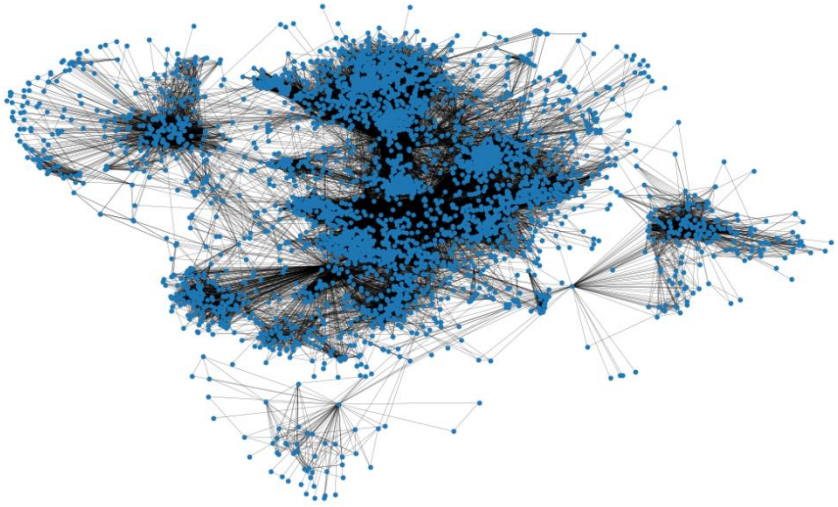
Each topic captures a unique facet of the discourse surrounding COVID-19 ranging from policy and health measures to social implications and community reactions.

2. Facebook Graph Analysis

Graph Analysis is an analysis of network which helps in understanding the structure and relationship of social networks like Facebook (Campbell *et al.*, 2013). It provides important information such as who is the most influential person/people in the network, define characteristics of groups of users, prediction of suitable things for users and so on.

For this analysis, social circles data from Facebook will be used which consists of 4039 nodes and 88234 edges ("SNAP: Network datasets: Social circles", n.d.).

Figure 11: Facebook Graph Visualization



Graph Statistical Description

Nodes: 4039

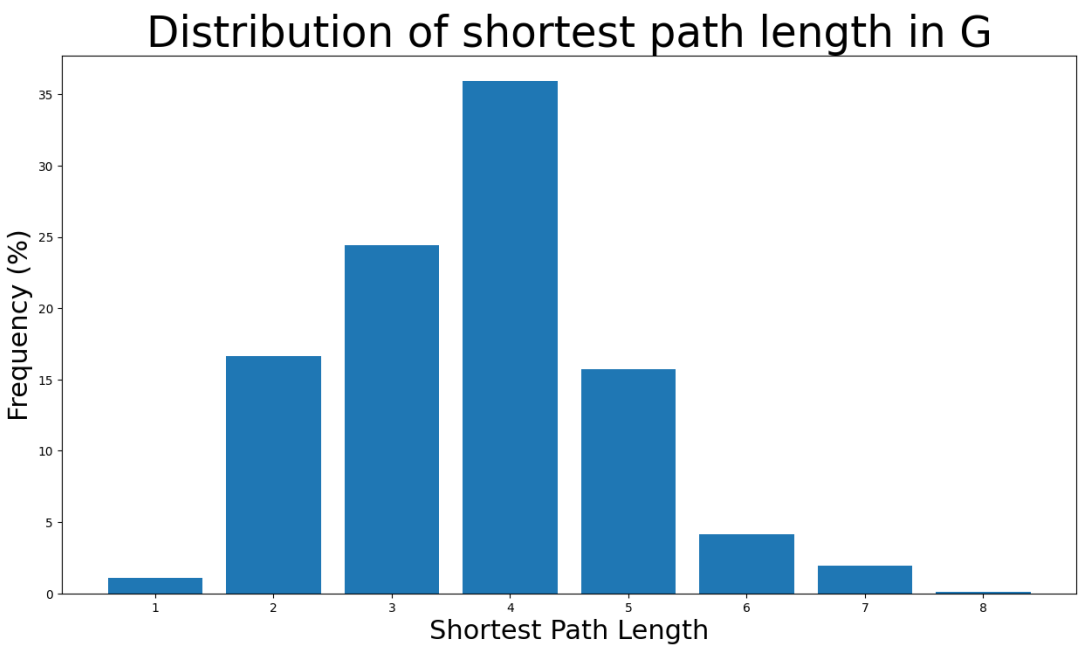
Edges: 88234

Average Node Degree: 43.69

Average Shortest Path Length for each node: 3.69

The above visualization shows a complex network graph of Facebook interactions highlighting the interconnected nature of user relationships. Nodes represent individual Facebook users, while edges depict the connections between them, indicating friendship or interaction. The average node degree of approximately 43.69 suggests that, on average, each user is connected directly to around 44 others, indicating a densely connected network.

Figure 12: Distribution of shortest path length in G



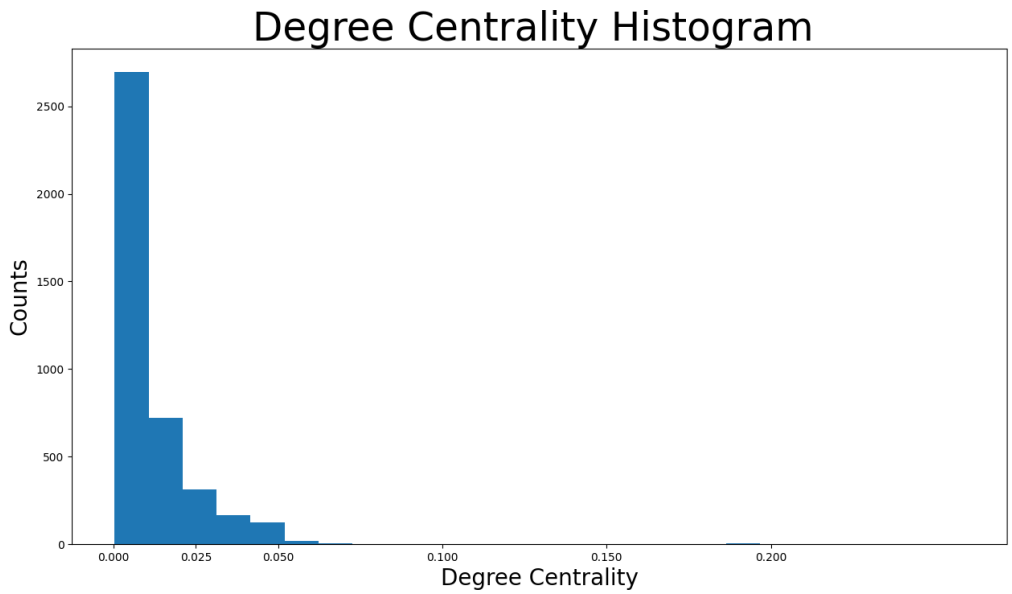
The bar chart illustrates the distribution of shortest path lengths within a network graph (G) showing that the most common shortest path length is 4, followed closely by lengths of 3 and 5 indicating a tightly interconnected network where most nodes can be reached from others within a few steps.

2.1 Find the most important nodes(individuals) in the network based on different Centrality Measures

Centrality Measures are used in network analysis to determine the most important nodes within a network by analyzing their structural properties, network patterns and behaviors (Grando *et al.*, 2016). Based on the network graph, different centrality measures such as degree centrality, betweenness centrality, closeness centrality and eigenvector centrality will be performed.

A. Degree Centrality: The users with highest degree centralities from the size of their nodes

Figure 13: Degree Centrality Histogram - Distribution of degree centralities



Degree Centrality Scores for the Nodes

[(107, 0.258791480931154),

(1684, 0.1961367013372957),

(1912, 0.18697374938088163),

(3437, 0.13546310054482416),

(0, 0.08593363051015354),

(2543, 0.07280832095096582),

(2347, 0.07206537890044576),

(1888, 0.0629024269440317)]

The number of neighbors for the nodes with the highest degree centralities

[(107, 1045),

(1684, 792),

(1912, 755),

(3437, 547),

(0, 347),

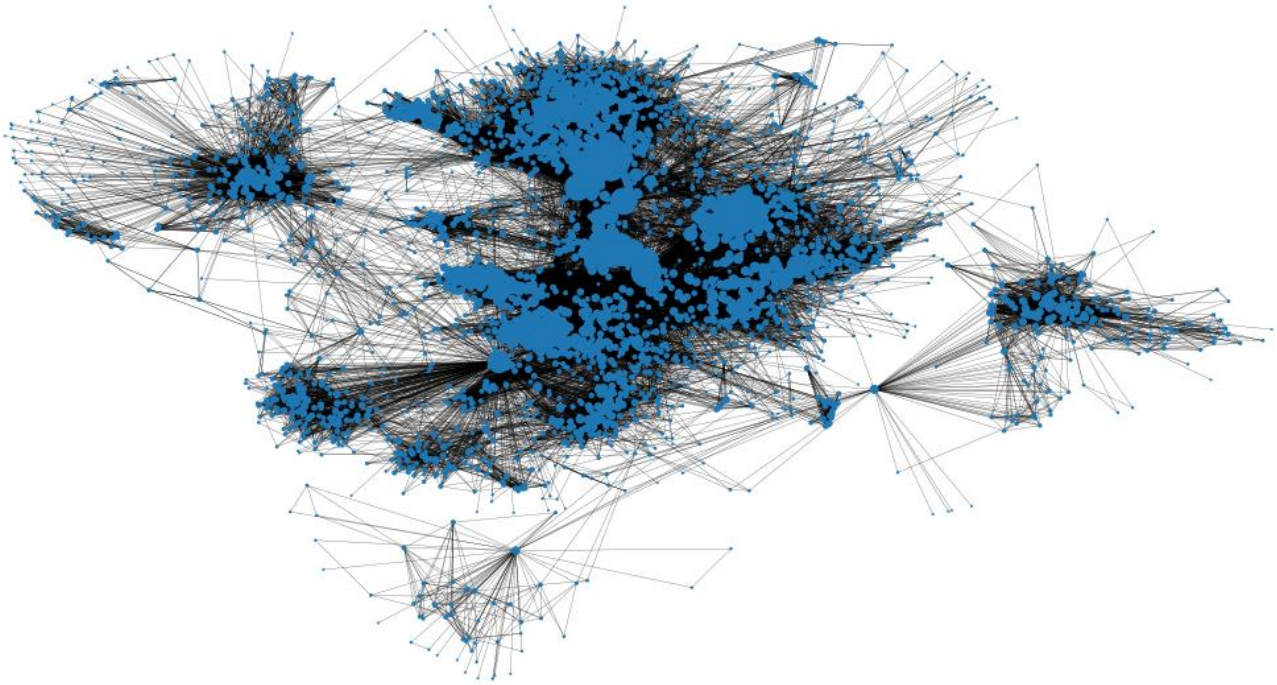
(2543, 294),

(2347, 291),

(1888, 254)]

Based on the above degree centrality scores, Node 107 has the highest centrality with a score of about 0.259 which means it is connected to approximately 25.9% of all possible nodes in the network. This suggests that node 107 is a very influential or important node within the network.

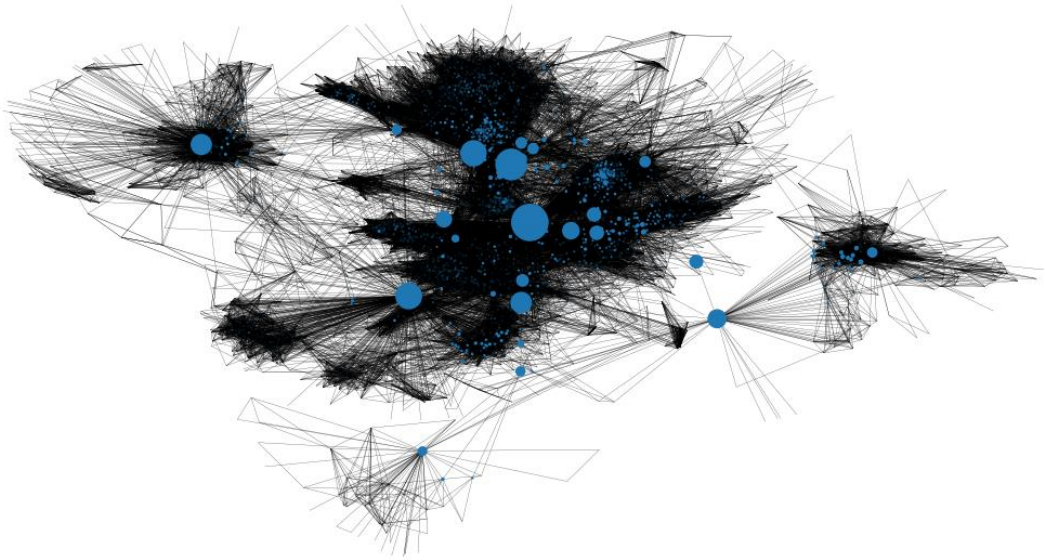
Figure 14: Highest Degree Centrality



The graph represents a highest degree centrality in a network that means nodes have more direct connections to other nodes within the network.

B. The nodes with the highest betweenness centralities

Figure 15: Highest Betweenness Centrality



Betweenness Centrality Scores for the Nodes

[(107, 0.4805180785560152),

(1684, 0.3377974497301992),

(3437, 0.23611535735892905),

(1912, 0.2292953395868782),

(1085, 0.14901509211665306),

(0, 0.14630592147442917),

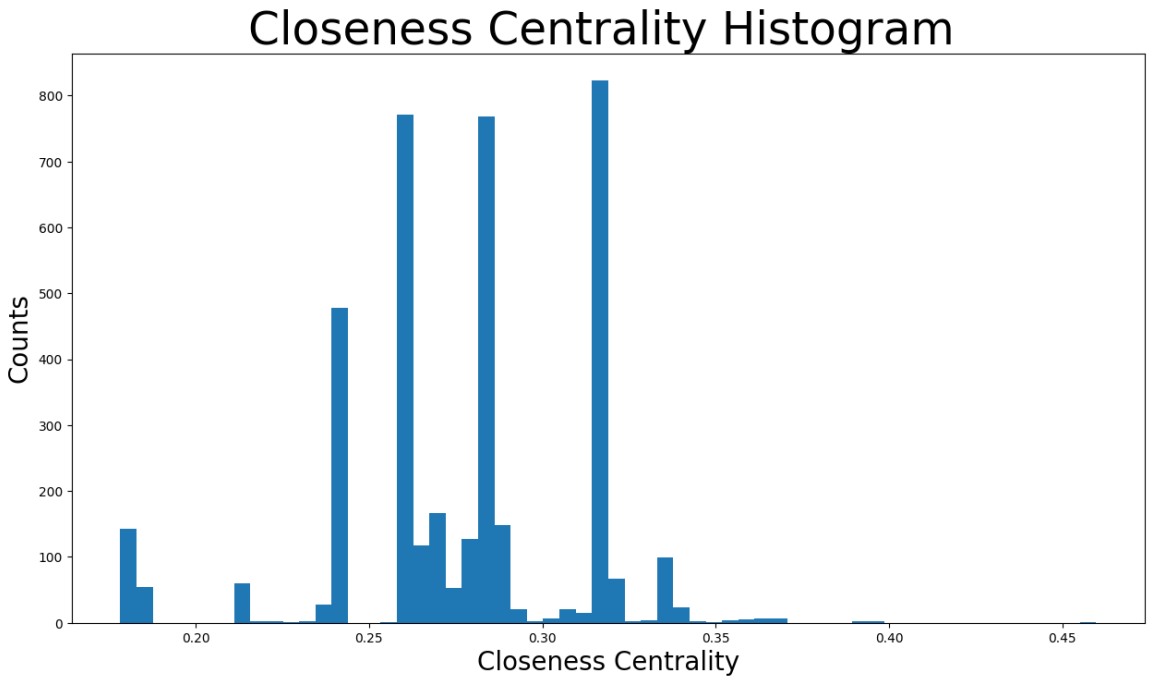
(698, 0.11533045020560802),

(567, 0.09631033121856215)]

In the above figure, the nodes with the highest betweenness centralities are marked with larger blue circles. For instance, node 107 has a betweenness centrality score of approximately 0.481 which tells that it plays a crucial role in connecting disparate parts of the network. The higher the betweenness centrality score, the more pivotal the node is in controlling the flow of information across the network.

C. The closeness centralities are distributed

Figure 16: Distribution of Closeness Centrality



Closeness Centrality Scores for the Nodes

[(107, 0.45969945355191255),

(58, 0.3974018305284913),

(428, 0.3948371956585509),

(563, 0.3939127889961955),

(1684, 0.39360561458231796),

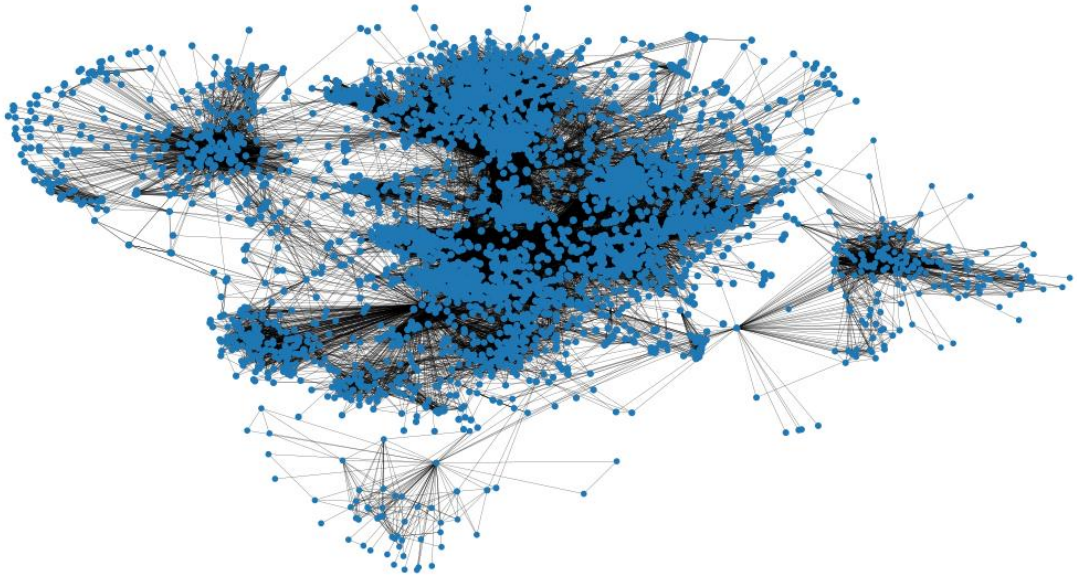
(171, 0.37049270575282134),

(348, 0.36991572004397216),

(483, 0.3698479575013739)]

The average distance of a particular node to any other node: 2.17

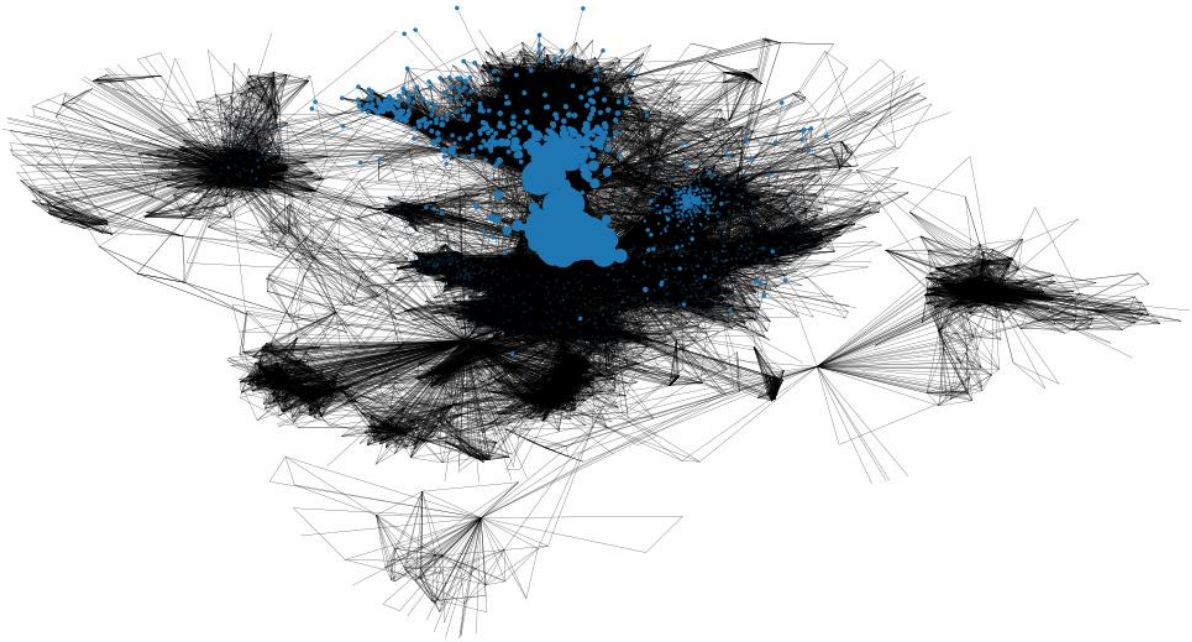
Figure 17: Closeness Centrality



The above figure shows the closeness centrality in which larger nodes indicate individuals with higher closeness centrality that implies they are closer to all other nodes in the network. With an average distance of 2.17 to any other node, nodes with higher closeness centrality play a pivotal role in enhancing connectivity and information exchange within the network.

D. The eigenvector centralities of nodes based on their size in the following representation

Figure 18: Eigenvector Centrality



Eigenvector Centrality Scores for the Nodes

[(1912, 0.09540696149067629),

(2266, 0.08698327767886552),

(2206, 0.08605239270584342),

(2233, 0.08517340912756598),

(2464, 0.08427877475676092),

(2142, 0.08419311897991795),

(2218, 0.0841557356805503),

(2078, 0.08413617041724977),

(2123, 0.08367141238206224),

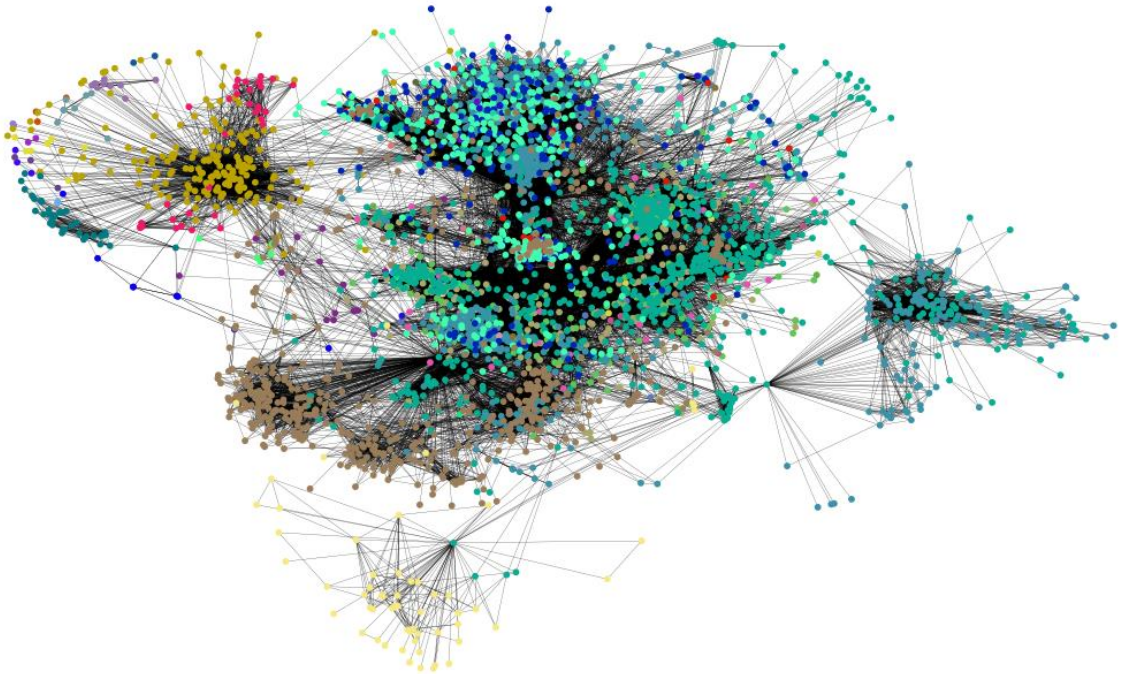
(1993, 0.0835324284081597)]

In the above figure, the eigenvector centralities of nodes are illustrated. Each node's eigenvector centrality score is crucial in determining its influence within the network, with higher scores indicating greater importance. For instance, node 1912 possesses an eigenvector centrality score of approximately 0.095 suggesting its significant impact on the network

2.2 Community Detection Algorithm

A community detection algorithm is a computational technique that is used to identify cohesive groups or communities within complex networks (Lancichinetti and Fortunato, 2009) These algorithms are particularly popular in network analysis across various domains such as social networks, biological networks and computer science disciplines (Fortunato, 2010).

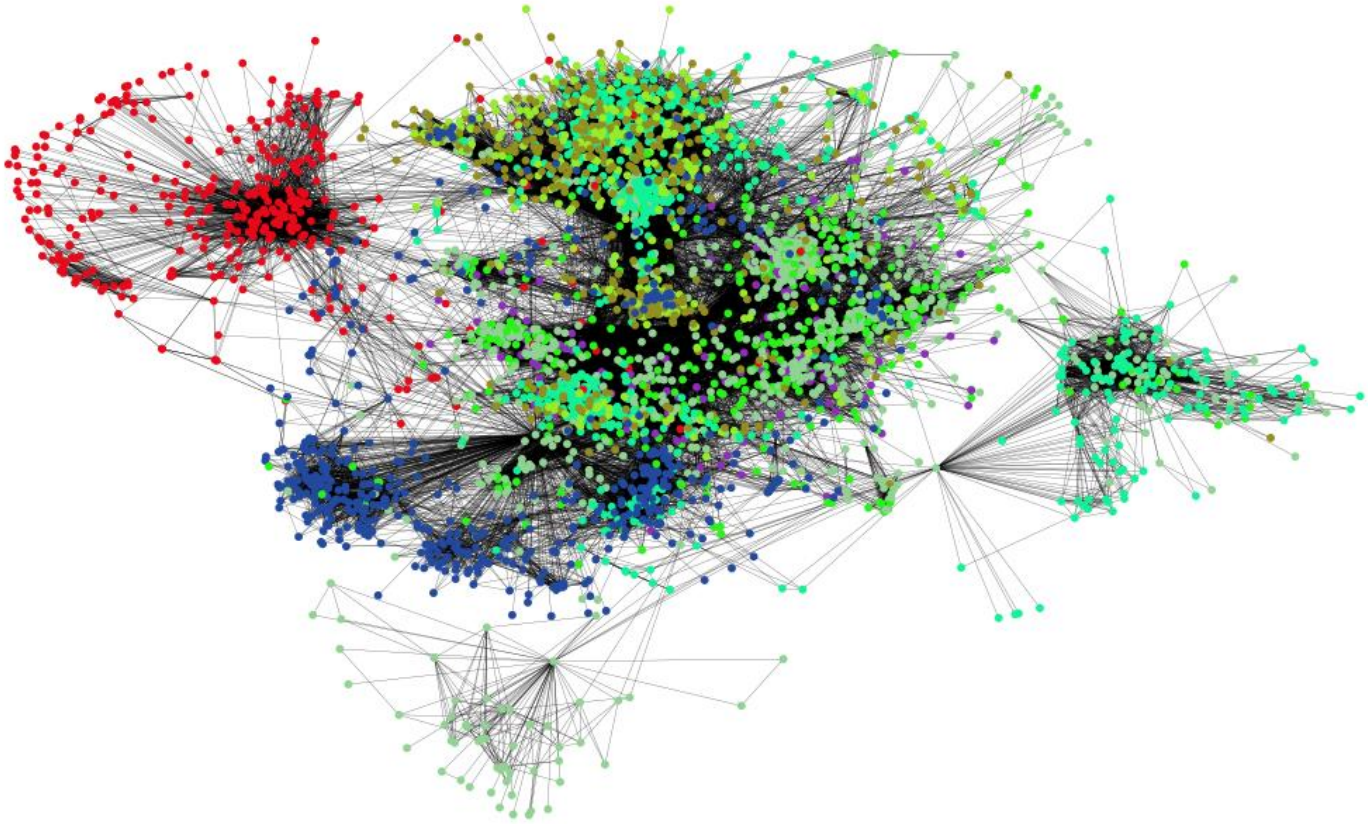
Figure 19: Community Detection Graph



Using community detection algorithms, total of 44 communities were detected.

Now, detecting only 8 communities and showing them in different color for each community nodes.

Figure 20: Selected 8 Communities are shown

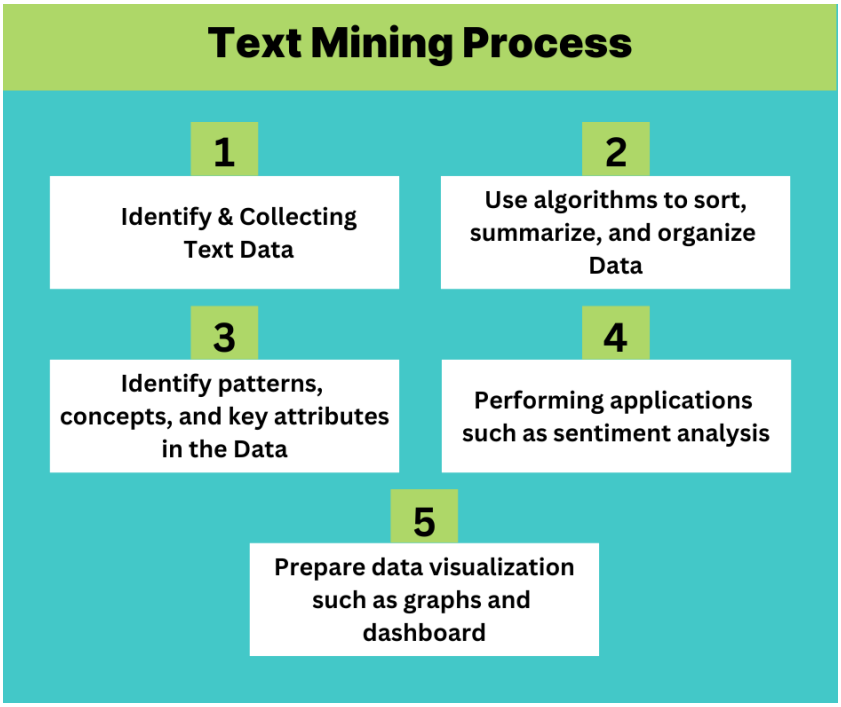


The above graph shows a complex network segmented into eight distinct communities, each represented by a unique color. The color-coding effectively highlights the structure of the network which shows how nodes within the same community are more densely connected to each other.

Part B: Text Mining

Text Mining or Text Analysis is a process of identifying textual patterns, trends, and insights from unstructured text data through the use of statistics, machine learning and linguistics (Younis, 2015). It can be applied to text-based dataset that are collected from social media, surveys website, forum posts, news article website, call transcripts, and so on. Text Mining has become more popular and practical for data analyst and scientist due to the development of big data technologies and deep learning algorithms that can analyze and provide meaningful results and information from vast amount of unstructured data (Liang and Dai, 2013).

Figure 21: The Text Mining Process (5 Steps)



3. Event/Campaign that happened in the UK or Worldwide recently (i.e. Brexit)

For this analysis, the chosen event is Brexit which means the withdrawal of the United Kingdom from the European Union as explained by (Sampson, 2017) in his journal of “Brexit: The Economics of International Disintegration”.

So, we will use twitter data and do data cleaning, preprocessing and filtering required data for our analysis.

Dataset Used: covid19_tweets.csv

3.1 Tweets Sentiment Distribution for Posts Related to Brexit

Figure 22: Distribution of Sentiments in Bar Diagram

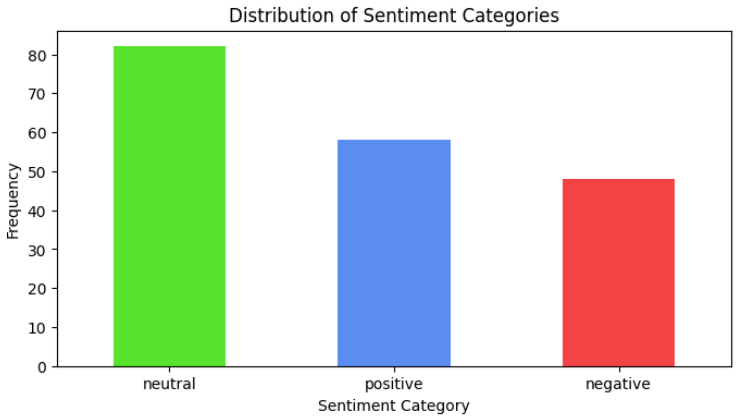
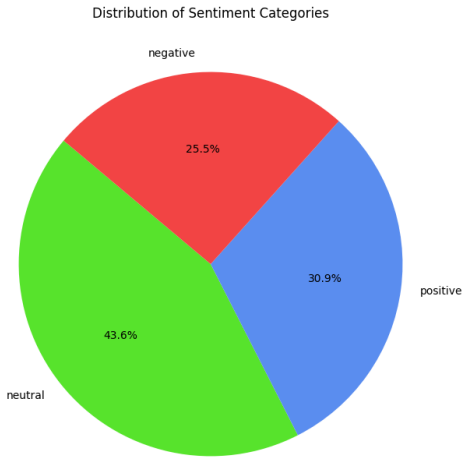


Figure 23: Distribution of Sentiments in Pie Chart



The above figures show the sentiment distribution of the tweets having hashtag #brexit. Neutral and Positive sentiments of the tweets are the highest with 43.6% and 30.9% respectively.

Figure 24: Positive Sentiment Word Cloud for #brexit tweets

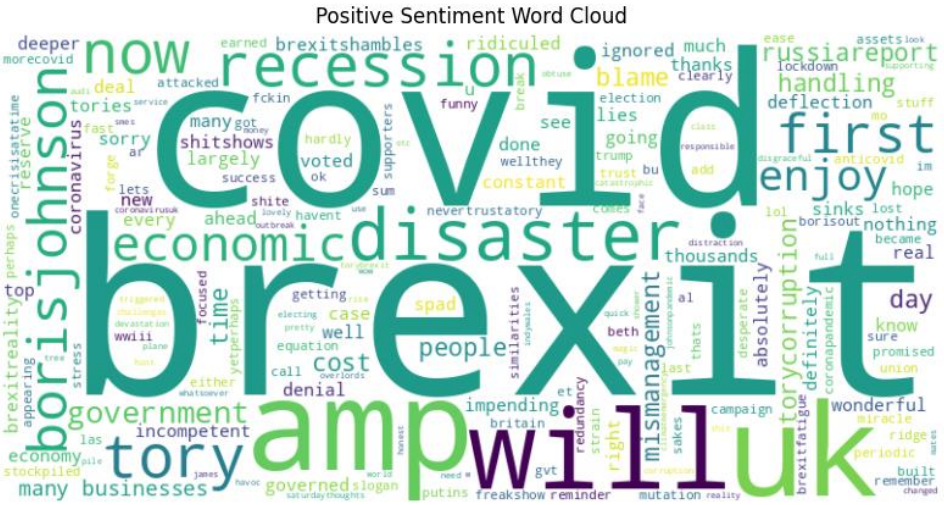


Figure 25: Negative Sentiment Word Cloud for #brexit tweets

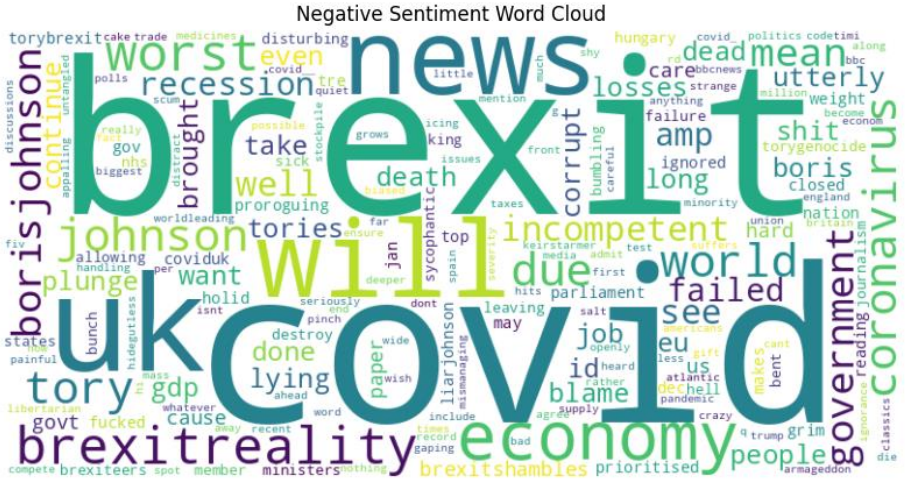
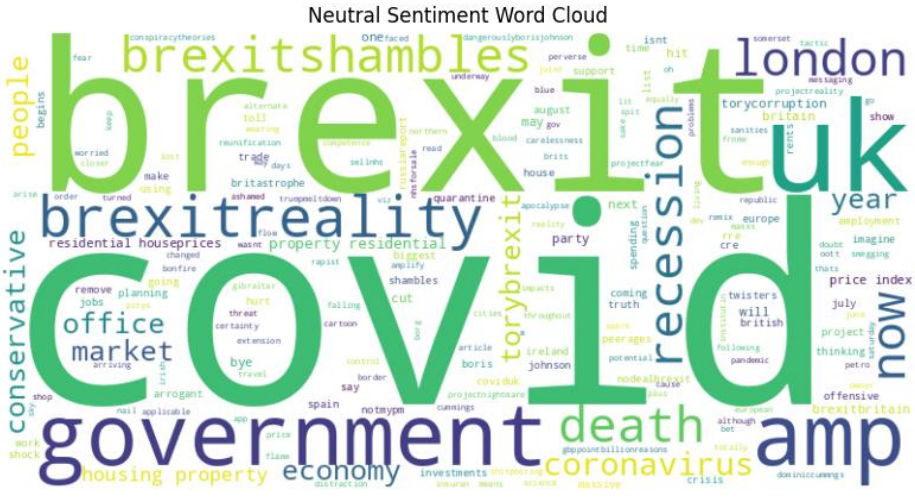


Figure 26: Neutral Sentiment Word Cloud for #brexit Tweets



Descriptive Statistics for Polarity Scores:

Mean Polarity: 0.008

Median Polarity: 0.000

Mode Polarity: 0.000

Standard Deviation of Polarity: 0.274

Descriptive Statistics for Subjectivity Scores:

Mean Subjectivity: 0.355

Median Subjectivity: 0.333

Mode Subjectivity: 0.000

Standard Deviation of Subjectivity: 0.330

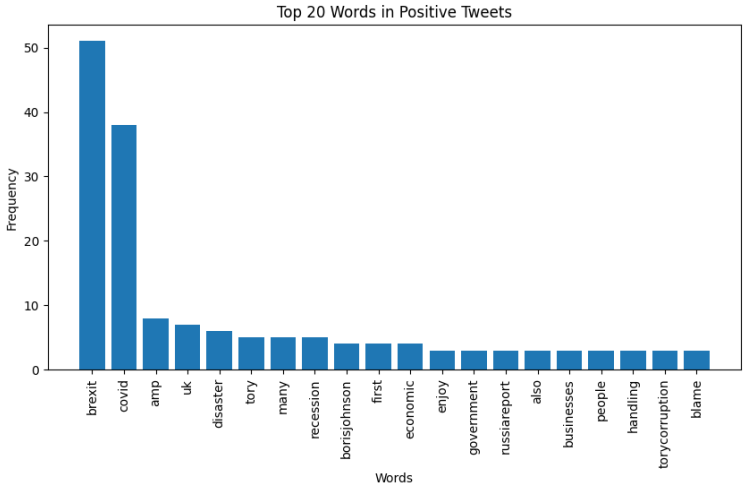
Counts of Each Sentiment Category:

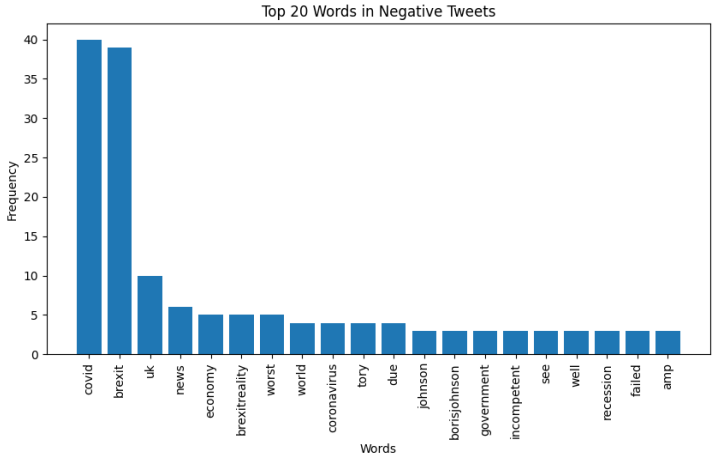
neutral 82

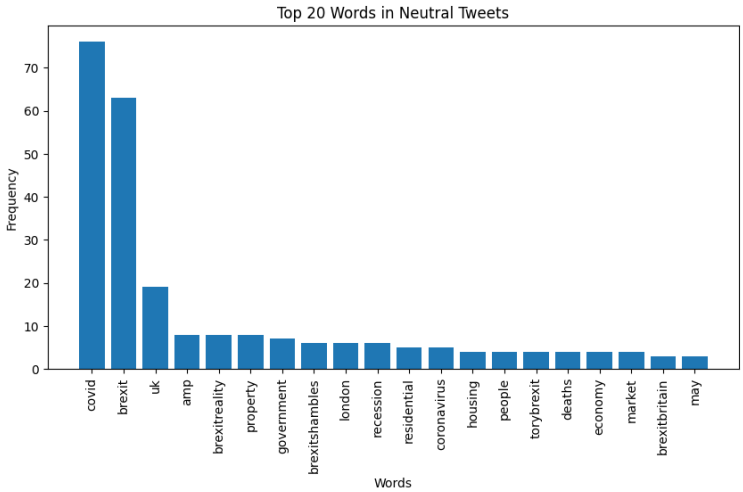
positive 58

negative 48

3.2 Frequency Distribution of Words in Sentiments

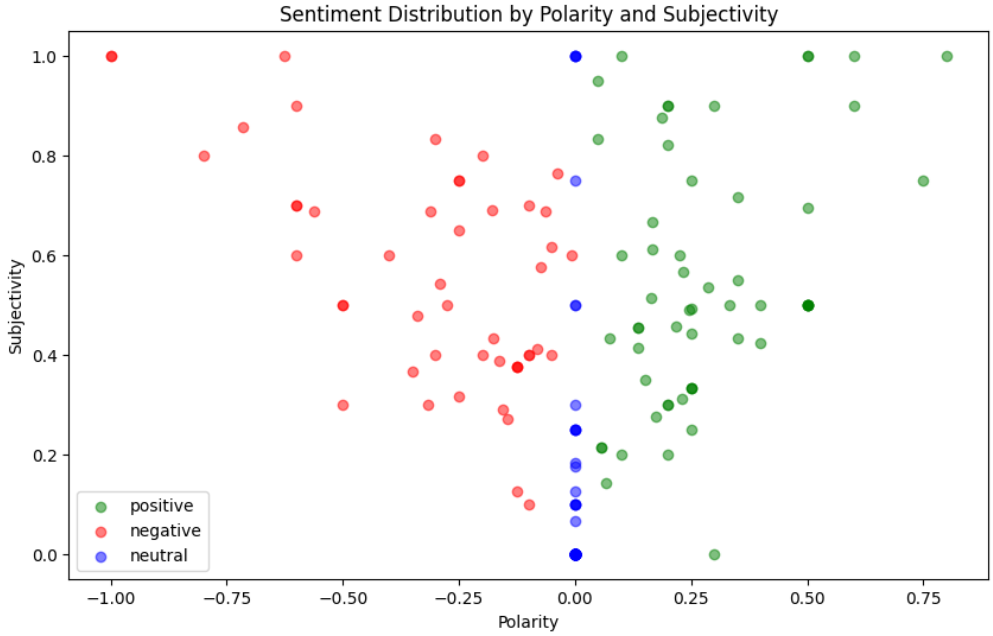






3.3 Sentiment Distribution by Polarity and Subjectivity

Figure 27:Sentiment Distribution Polarity



4. News Article Analysis (Using APIs)

In the News Article Analysis, we use the News API offered by newsapi.org ("Documentation - News API", n.d.) to perform different types of analysis based on news article. In addition, we will also implement machine learning technique to predict news category based on the title of the news.

4.1 Cleaning and Preprocessing on the Articles

To ensure the data from the articles is suitable for analysis, we apply a sequence of preprocessing steps to the text extracted from the API as follows:

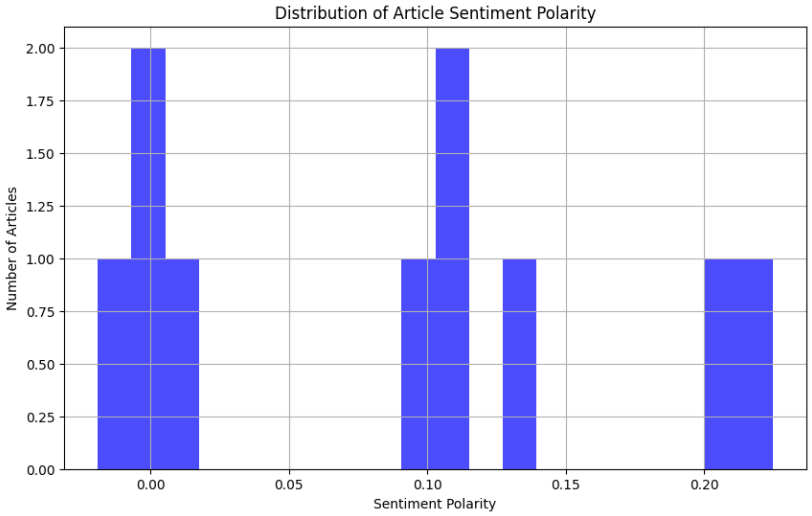
a. Tokenization: We begin by breaking down the text into individual words or tokens.
b. Lowercasing: Next, all tokens are converted to lowercase.
c. Stop-word Removal: We then remove common words that typically add little analytical value.
d. Lemmatization: Finally, words are reduced to their base or dictionary form.

4.2 Descriptive Analysis of the collected articles

In this, we focus on quantitatively summarizing the features of our dataset through various statistical measures to understand the underlying patterns in the news articles.

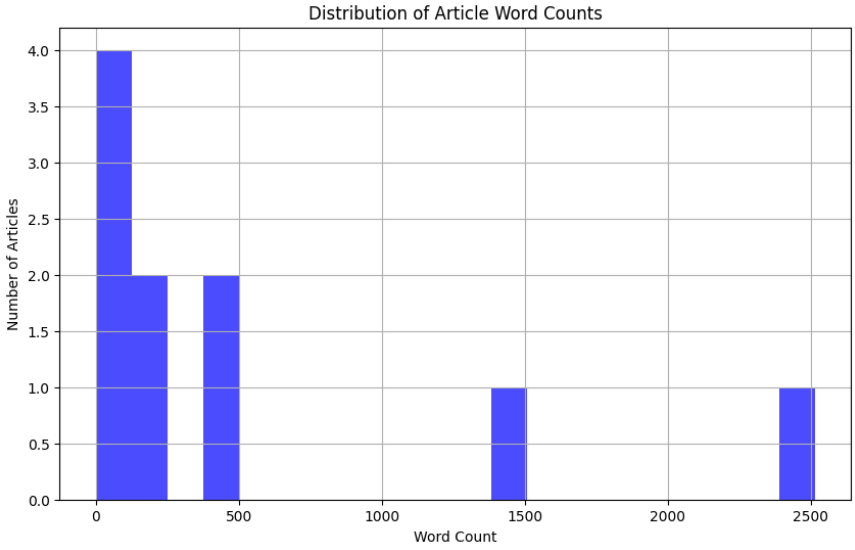
A. Distribution of Article Sentiment Polarity

Figure 28: Distribution of Articles by Sentiment Polarity



B. Distribution of Word Count

Figure 29: Distribution of Word Counts



4.3 LDA Topic Modeling Techniques to discover key topics

Figure 31: Topic Modeling to discover topics

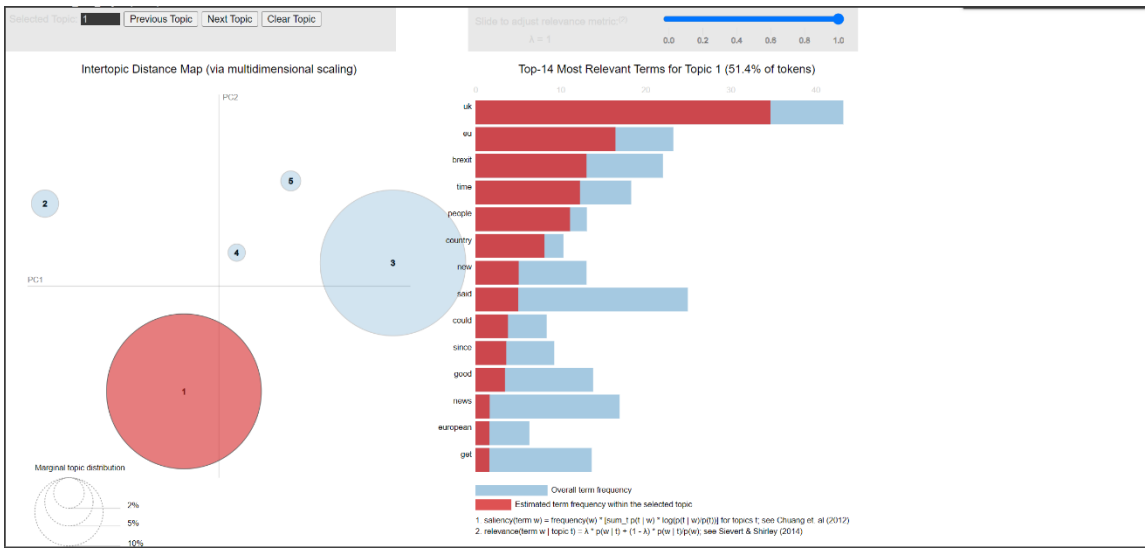


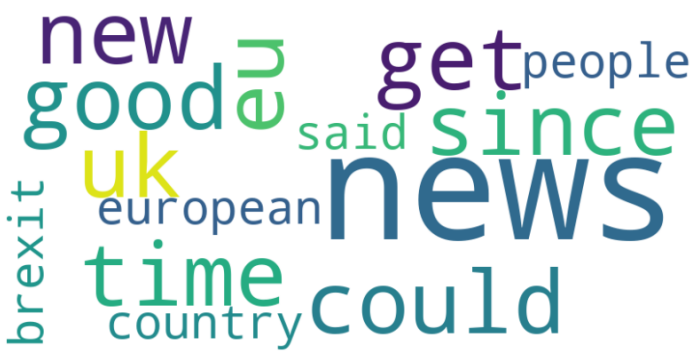
Figure 32: Word Cloud for Topic 1

Figure 33: Word Cloud for Topic 2

Figure 34: Word Cloud for Topic 3

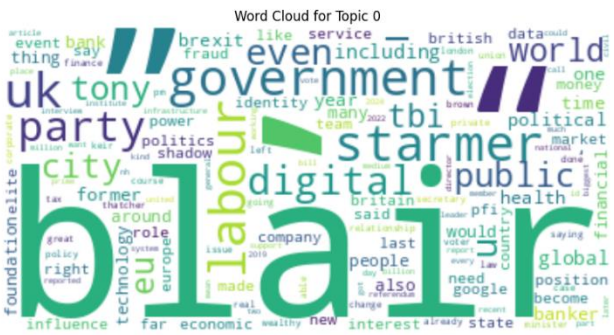
Figure 35: Word Cloud for Topic 4

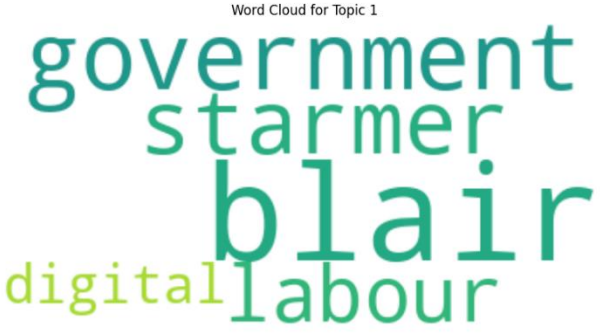
Figure 36: Word Cloud for Topic 5



4.4 LSA Topic Modeling techniques to discover key topics

Based on the implementation of LSA topic modeling techniques, word clouds are generated for the topics which are given below.





4.5 Summary of one of the news articles

In this, one lengthy article from the lists of articles is chosen and summarized using the weighted frequency of occurrences and scoring sentences based by weighted frequencies of words. Thus, selecting only the top sentences with the highest score will give us a summarized text from that article.

Original Article (3933 Words)

Figure 37: Original Article Text

Summarized Article (333 words):

A recent feature article in the London Times, titled "Tony Blair: Politics Is for the Weird and the Wealthy", provides a glimpse of just how much influence Blair and TBI are likely to wield during a Starmer government:

Starmer, who shared a stage with Blair at the TBI's Future of Britain conference last summer, has populated his team with Blairites — including the former Blair special adviser Matthew Doyle, now Starmer's director of communications; the former Blair strategist and speechwriter Peter Hyman, who is a senior adviser; and another former Blair special adviser, Peter Kyle, now the shadow science secretary. TBI was spawned in 2017 by rolling together all of Blair's for-profit and non-profit ventures, including the Tony Blair Faith Foundation, the Tony Blair Sports Foundation, the Tony Blair Governance Initiative, and his consulting firm Tony Blair Associates, into one vehicle. Starmer is favourite to win not because of a groundswell of support for his vision or candidacy — the UK public view the party under Starmer even less favourably than under Ed Miliband — but because support for the governing (if you can call it that) Conservative Party is in freefall:

Remarkable data, the UK public actually view the Labour Party under Keir Starmer's as far less capable than it was under the tepid Ed Miliband, and yet the useless plank Starmer is destined to become PM entirely thanks to the Tory's self-destruction. In the comments thread to a recent post of mine, Colonel Smithers, a regular NC commentator who is UK power politics adjacent, suggested that the next Labour government will be even more in thrall to the City of London than Blair or Brown's:

From late November, I have become involved with the trade body representing foreign banks operating in the City, including their engagement with Labour. It also goes without saying that issues around digital technology (digital vaccine certificates digital ID, CBDCs, biometric identifiers, digital censorship, digital health data…) will feature heavily in a Starmer / Blair 2.0 government.

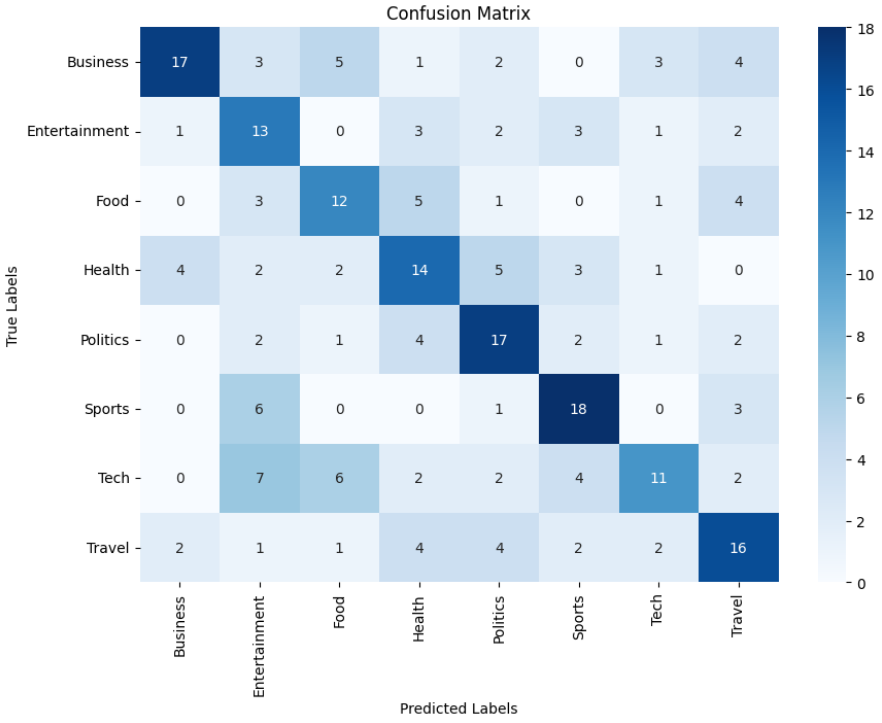
5. Machine Learning and Deep Learning Implementation

In this section, we use machine learning technique logistic regression to predict news categories whereas TensorFlow and Keras for sentiment analysis of tweets. This approach leverages the strengths of both machine learning and deep learning to handle different types of data and tasks.

5.1 News Article Category Prediction using Logistic Regression

Here, we use logistic regression to classify news articles into different categories. This method involves using the Python libraries such as "NewsApiClient" for fetching news data, "pandas" for data manipulation, "nltk" for text processing like tokenization and removing stopwords, and "sklearn" for creating a machine learning pipeline.

Figure 38: Confusion Matrix



The above confusion matrix visualizes the performance of a model that predicts categories for news articles such as Business, Entertainment, Food, Health, Politics, Sports, Tech, and Travel. For instance, 17 Business articles were correctly predicted as Business, but 3 were incorrectly predicted as Entertainment. The matrix highlights the model's accuracy and its errors, showing which categories are often confused with others, like Tech articles frequently being misclassified as

Business. This helps in identifying strengths and weaknesses in the classification model.

Testing the Model (Outcome)

Input: “New Vegan Restaurant Chain Expands Across the United States”

Predicted Category: Entertainment

Input: “Oscar Winners Announced: Surprising Wins in Major Categories”

Predicted Category: Politics

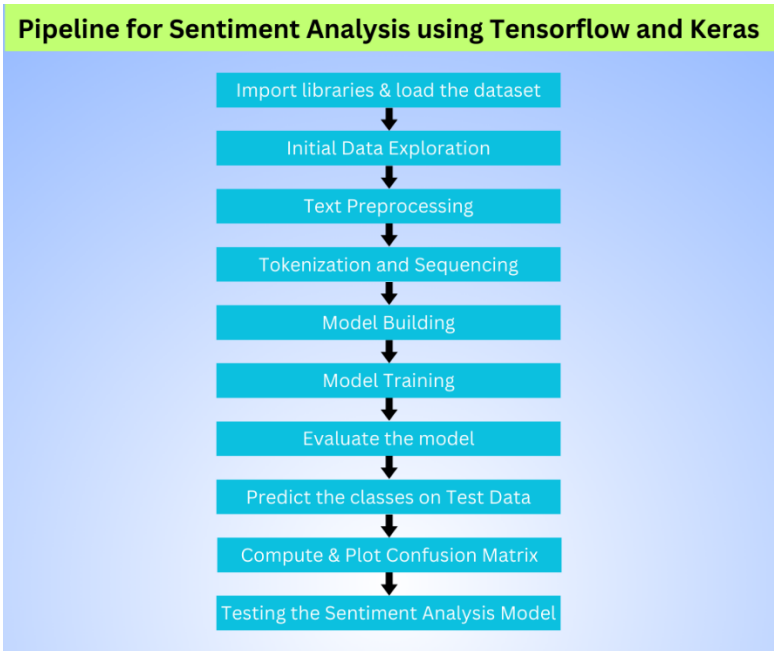
Accuracy is: 0.49166666666666664

5.2 Sentiment Analysis on Tweets using TensorFlow and Keras

Here, we perform sentiment analysis process where machine learning frameworks TensorFlow and Keras are used to analyze the sentiment of tweets. This involves classifying tweets into categories such as positive, negative, or neutral based on their content, utilizing deep learning models to process and predict sentiments efficiently. on tweets using TensorFlow and Keras.

Dataset Used: tweets_sentiment.csv

Figure 39: Sentiment Analysis on Tweets using TensorFlow and Keras

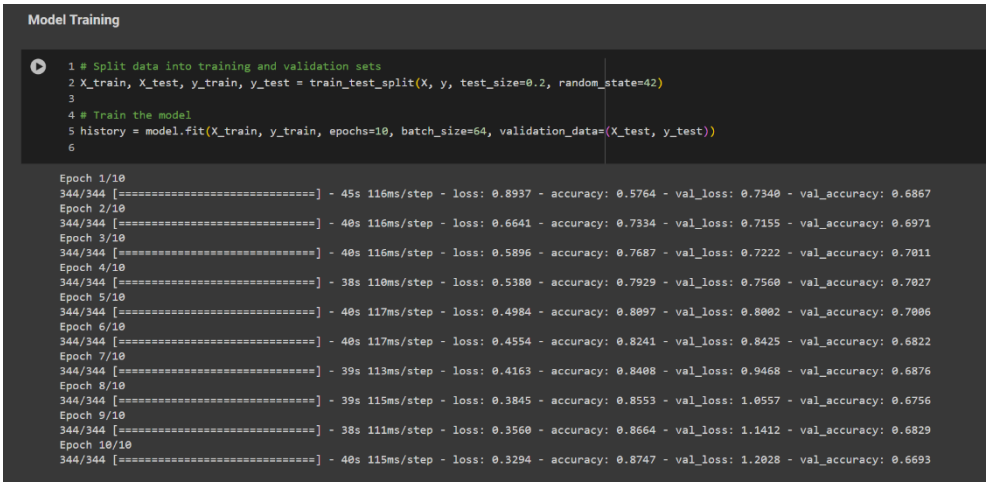


The above image describes the pipeline or processes used for sentiment analysis using Tensorflow and Keras.

The process begins with importing necessary libraries and loading the dataset, followed by initial data exploration to understand the dataset's characteristics. Next, text preprocessing is applied to clean and prepare the data. This includes tokenization and sequencing to convert text into a format suitable for model input. The subsequent steps involve building and training the machine learning model, evaluating its performance, and making predictions on test data. Finally, the results are analyzed using a confusion matrix, and the model is tested to confirm its effectiveness in classifying sentiments.

Model Training

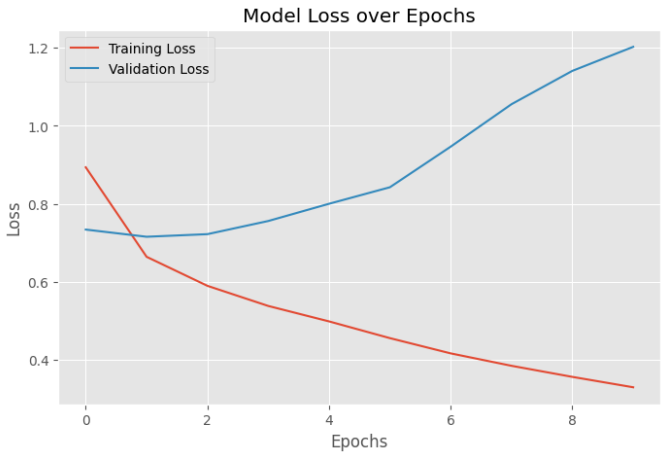
Figure 40: Model Training after splitting data into training and test set



The training results indicate that while the model's performance on the training set consistently improved over the 10 epochs which is evidenced by decreasing loss and increasing accuracy. The best accuracy achieved by the model on the validation set was 0.7027, which occurred in the fourth epoch.

Evaluate the Model (Model Loss & Model Accuracy)

Figure 41: Model Loss Over Epochs

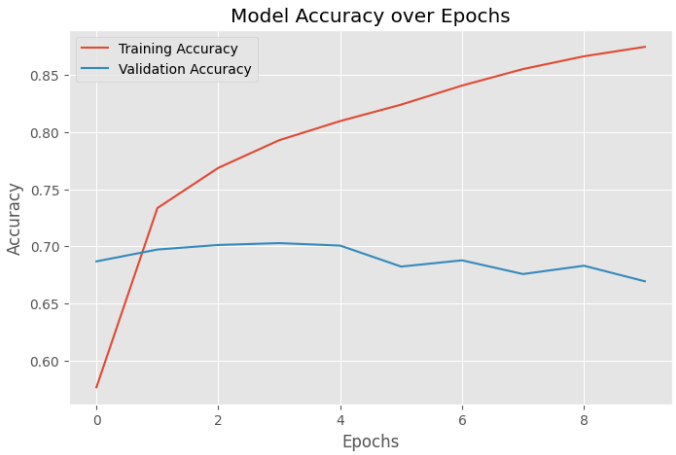


This graph displays the training and validation loss:

Training Loss (Red Line): The training loss decreases significantly from around 1.2 to below 0.4 over the epochs. This reduction in loss indicates that the model is becoming increasingly proficient at predicting the training data.

Validation Loss (Blue Line): Conversely, the validation loss begins at about the same level as the training loss but increases over time, ending around 1.1. This trend indicates that the model, while fitting the training data better over time, is performing worse when exposed to new, unseen data.

Figure 42: Model Accuracy Over Epochs



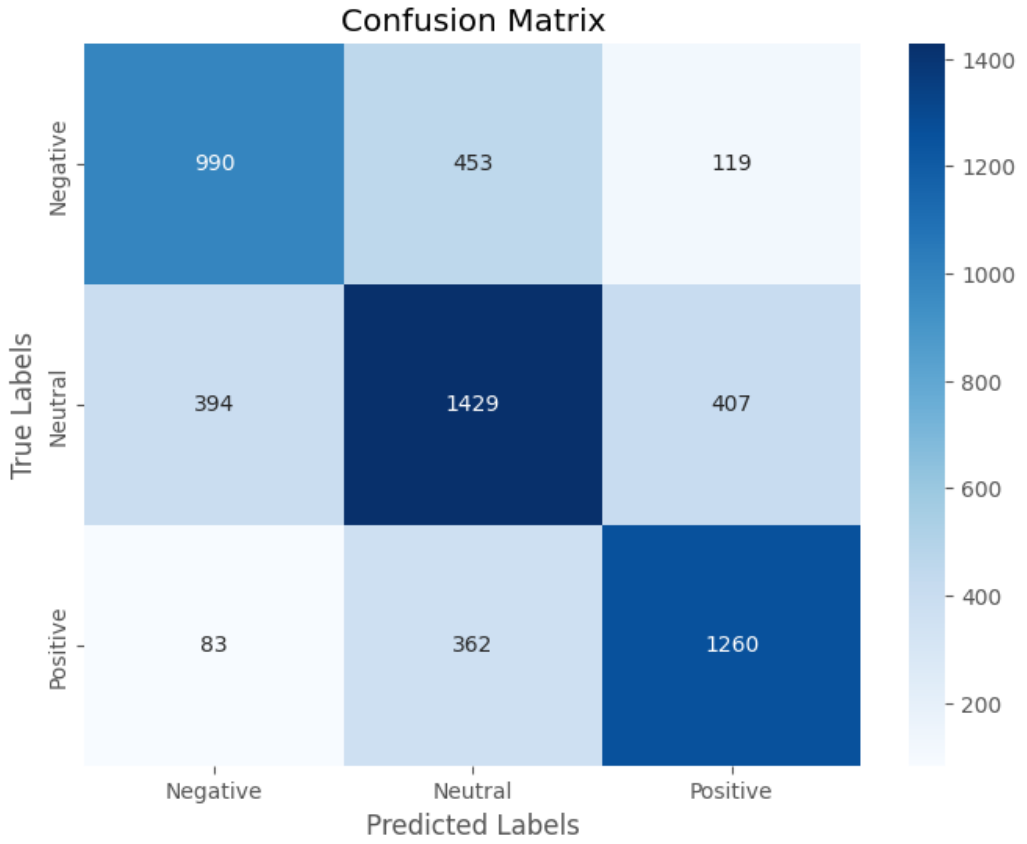
This graph displays two curves: training accuracy and validation accuracy:

Training Accuracy (Red Line): This line indicates that the model's accuracy on the training dataset improves steadily over epochs, starting from around 65% and approaching close to 90% by epoch 9. This suggests that the model is learning effectively from the training data.

Validation Accuracy (Blue Line): Contrarily, the validation accuracy starts around 66% and only marginally increases to about 67% by epoch 9. This relatively flat line suggests that while the model performs increasingly well on the training data, its performance on unseen validation data does not improve significantly.

Confusion Matrix

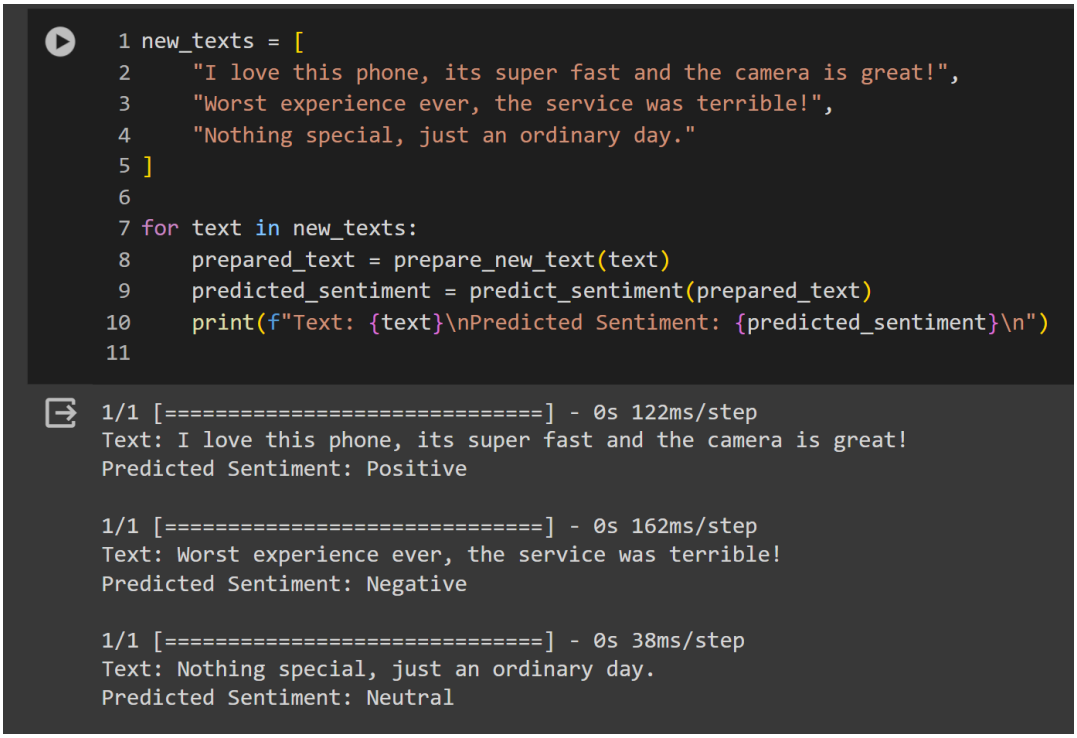
Figure 43: Confusion Matrix



The above image is a confusion matrix which is a tool used to evaluate the performance of a machine learning model particularly for classification tasks. It shows the counts of true and predicted classifications across three categories: negative, neutral, and positive. The diagonal entries (990 for negative, 1429 for neutral, 1260 for positive) indicate correct predictions, demonstrating the model's effectiveness in identifying each sentiment correctly.

Testing with Examples

Figure 44: Testing Sentiment Model with sample tweets.



```
new_texts = [
    "I love this phone, its super fast and the camera is great!",
    "Worst experience ever, the service was terrible!",
    "Nothing special, just an ordinary day."
]

for text in new_texts:
    prepared_text = prepare_new_text(text)
    predicted_sentiment = predict_sentiment(prepared_text)
    print(f"Text: {text}\nPredicted Sentiment: {predicted_sentiment}\n")

```

```
1/1 [==============================] - 0s 122ms/step
Text: I love this phone, its super fast and the camera is great!
Predicted Sentiment: Positive

1/1 [==============================] - 0s 162ms/step
Text: Worst experience ever, the service was terrible!
Predicted Sentiment: Negative

1/1 [==============================] - 0s 38ms/step
Text: Nothing special, just an ordinary day.
Predicted Sentiment: Neutral
```

The above image demonstrates the model's ability to classify sentiments accurately based on the content of the texts.

6. References

Annamoradnejad, I. and Habibi, J. (2019), “A Comprehensive Analysis of Twitter Trending Topics”, *2019 5th International Conference on Web Research, ICWR 2019*, Institute of Electrical and Electronics Engineers Inc., pp. 22–27, doi: 10.1109/ICWR.2019.8765252.

Campbell, W.M., Dagli, C.K. and Weinstein, C.J. (2013), “SOCIAL NETWORK ANALYSIS WITH CONTENT AND GRAPHS”, *LINCOLN LABORATORY JOURNAL ν VOLUME*, Vol. 20.

“Documentation - News API”. (n.d.). , available at: https://newsapi.org/docs (accessed 3 May 2024).

Doshi, Z., Nadkarni, S., Ajmera, K. and Shah, N. (2017), “TweerAnalyzer: Twitter Trend Detection and Visualization”, *2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017*, Institute of Electrical and Electronics Engineers Inc., doi: 10.1109/ICCUBEA.2017.8463951.

Fortunato, S. (2010), “Community detection in graphs”, *Physics Reports*, North-Holland, Vol. 486 No. 3–5, pp. 75–174, doi: 10.1016/J.PHYSREP.2009.11.002.

Grando, F., Noble, D. and Lamb, L.C. (2016), “An analysis of centrality measures for complex and social networks”, *Proceedings - IEEE Global Communications Conference, GLOBECOM*, doi: 10.1109/GLOCOM.2016.7841580.

Lancichinetti, A. and Fortunato, S. (2009), “Community detection algorithms: A comparative analysis”, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, American Physical Society, Vol. 80 No. 5, p. 056117, doi: 10.1103/PHYSREVE.80.056117/FIGURES/8/MEDIUM.

Liang, P.W. and Dai, B.R. (2013), “Opinion mining on social media data”, *Proceedings - IEEE International Conference on Mobile Data Management*, Vol. 2, pp. 91–96, doi: 10.1109/MDM.2013.73.

Sampson, T. (2017), “Brexit: The Economics of International Disintegration”, *Journal of Economic Perspectives*, American Economic Association, Vol. 31 No. 4, pp. 163–84, doi: 10.1257/JEP.31.4.163.

Sapountzi, A. and Psannis, K.E. (2018), “Social networking data analysis tools & challenges”, *Future Generation Computer Systems*, North-Holland, Vol. 86, pp. 893–913, doi: 10.1016/J.FUTURE.2016.10.019.

“SNAP: Network datasets: Social circles”. (n.d.). , available at: https://snap.stanford.edu/data/ego-Facebook.html (accessed 8 May 2024).

Tang, X. and Yang, C.C. (2012), “TUT: A statistical model for detecting trends, topics and user interests in social media”, *ACM International Conference Proceeding Series*, pp. 972–981, doi: 10.1145/2396761.2396884.

Younis, E.M.G. (2015), “Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study”, *International Journal of Computer Applications*, Vol. 112 No. 5, pp. 975–8887.

7. Appendices

Source Code File Structures:

1. **Part A - Twitter Data Analysis.ipynb**

This file includes code for Twitter Trends Analysis. It includes topic modeling using LDA, word cloud visualization, sentiment analysis, devices used for tweets visualization, and trends analysis.

2. **Part A - Heatmap Visualization.ipynb**

This file includes code for Heatmap visualization and it is a part of Twitter Trend Analysis.

3. **Part A - Facebook Graph Data Analysis.ipynb**

This file includes graph data analysis, centrality measures and visualization, and community detection algorithm.

4. **Part B - TextMining for Brexit Event Analysis.ipynb**

This file includes "Brexit" event analysis such as sentiment analysis, word cloud visualization based on sentiment, statistical description such as frequency distribution, and sentiment polarity using tweets data.

5. **Part B - TextMining News Article Analysis Brexit.ipynb**

This file includes text mining of news articles such as descriptive analysis such as publication time distribution, word count distribution, article sentiment polarity, LDA topic modeling, word cloud based on LDA topic modeling, LSA topic modeling, word cloud for LSA topic modeling and summarization of one article.

6. **Part B - Machine Learning for News Category Prediction.ipynb**

This file includes implementation of Machine Learning technique Logistic Regression for News Category Prediction using NewsAPI data.

7. **Part B - Sentiment Analysis of Twitter Data Using Deep Learning (Tensorflow and Keras).ipynb**

This file includes sentiment analysis of twitter data using deep learning library such as TensorFlow and Keras.