Statisitcal, Network & Sentiment Analysis of Twitter and Newspaper Topic Modeling & Summarization

Hamza Waheed, ID: 21171402 [*] [1]

Word Count: 2369

1. Introduction

Social media has become the most essential part of today's worldwide web. Unlike traditional media, social media brings interaction, live two-way communication. This engagement has seen unprecedented growth in the usage of Social Media Platforms like Facebook, YouTube, Twitter, WhatsApp, LinkedIn, Reddit, Instagram etc.

Twitter is one of the leading Social Media platforms when it comes to user engagement. Twitter has an estimated 396.5 million monthly active users (Dean). The number of estimated live-posts called ”tweets” on Twitter are estimated to be 575k per minute (Ali).

Unlike other platforms, world leaders, musicians, politicians, businessman, celebrities all engage on Twitter. Some notable personalities on Twitter include: Elon Musk, Barack Obama, Cristiano Ronaldo, J.K Rowling, Boris Johnson etc. This endorsement from different fraternities of people gives Twitter an edge over other platforms.

The most notable feature, offered by Twitter, includes access to the major event widely discussed by the community of its users. This interaction produces a sudden increase in real time mentions of the topic as it unfolds. This feature is called ”Twitter Trend”. Twitter trends are based on popular topics & events discussed within a particular region or area. This region can include Worldwide and countrywide trends. These trends may range from Sports, festivals and Political events. The spike in real-time mentions produces trends that include specific terms used in that particular trend.

Many countries are combating the problem that arise from Real and Fake trends. Many authorities use these trends for their political gains. ”Real” & ”Fake News” is one of the most pressing issues faced by many countries. Many uprisings around the globe, are attributed to the targetted Twitter campaign (Daniel, 2013). Many governments, businesses and investment companies are focusing their capital to get the better insights out of these trends. Not only will it help to curtail any potential threat, but also better advertise social issues and product advertisements.

This study focuses on the following;

- Popular worldwide trends
- The number of tweets in those trends.
- Platform used for tweets
- Geographical relevance of tweets
- What sources to trust
- Relationship between tweets & likes
- Relationship between followers & likes

2. Twitter Analysis

2.1. Twitter Trends

This study focuses on the Top USA Twitter trend on 19/04/2022 at 10:00pm BST.

Figure 1, shows top trending topics in the USA. The hashtag ”Liverpool” trends at the top, as Liverpool FC won the match 4-0 against its arch rival Manchester United. We will perform exploratory analysis to get more insights behind the tweets involved with ”Liverpool” trends. The total number of tweets extracted were, 19600.

2.2. Devices Used

The fig 2 3 show the client platform used for tweets. It's evident from the graph that most tweets were done using the ”iPhone” followed by ”Android” and a small portion were done using ”Tweet bot” and ”Tweet deck”.

2.3. Verified Devices

Fig 4 highlights that only 4% of the devices used in tweets were verified. Meanwhile, a large majority of the account were unverified.

[1] M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: Hamza Waheed <Hamza.Waheed2@mail.bcu.ac.uk>.

Web Social Media Analytics and Visualization, CMP 7202 - Web Social Media Analytics and Visualization.

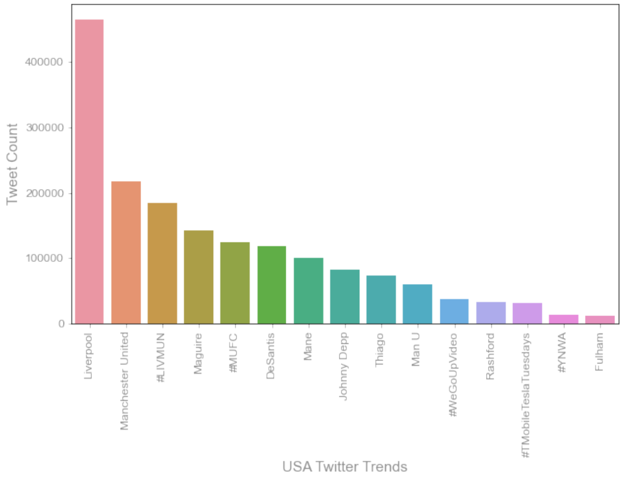


Figure 1. USA Twitter trends on 19/04/2022

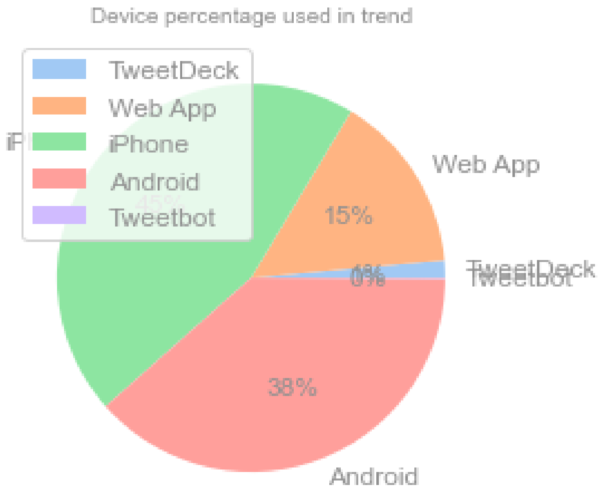


Figure 2. Percentage of device platforms used for Liverpool trend

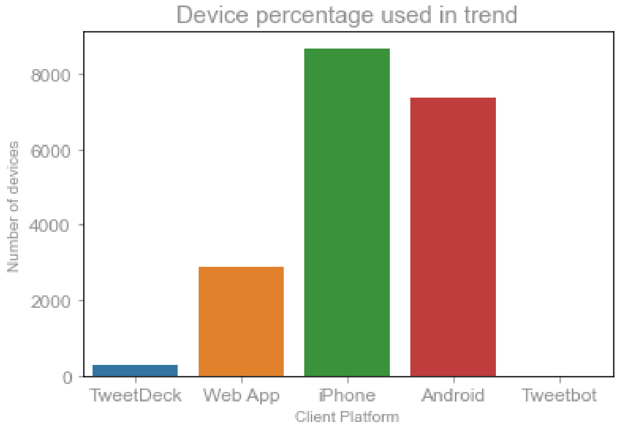


Figure 3. Number of devices used for Liverpool trend

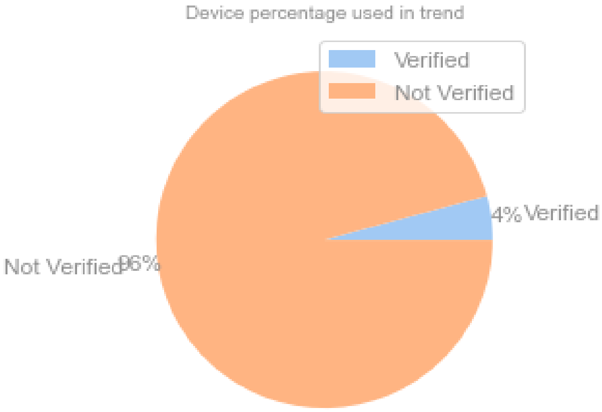


Figure 4. Percentage of verified devices used

2.4. Wordcloud & Frequency Distribution of Tweets

The fig 6, highlights ”Liverpool”, ”Manchester”, ”Cristiano” and ”United” as the most common topic of tweets. Fig 5 shows, the term ”Liverpool” was used more than, 16000 times. Also, the term ”United” & ”Manchester” had a frequency of 8000 & 6000 respectively.

2.5. Location Of Tweets

The location of the tweets were extracted using the Latitude, Longitude attributes in the Tweet object. ”Folium” was used to plot the location of tweets. Below are the location of tweets in ”Liverpool” trends.

Fig 7, 8, 9, 10 highlight the locations of the tweets. One key finding of the trend is the majority of the tweets involved in the trend ”Liverpool” were from the United Kingdom, where the match was taking place. Also, it’s worth noting that ”Folium” clusters similar locations into one group and aggregates location sum as a cluster for each location . At each zoom level, the clusters are further subdivided to reveal more locations until we reach the max zoom level, as shown in Fig 7, 8.

2.6. Retweet & Likes

Fig 11, highlights relation b/w Retweets and Likes. The relation between Retweets & likes is somewhat Linear. However, it’s worth noting that some tweets have very high

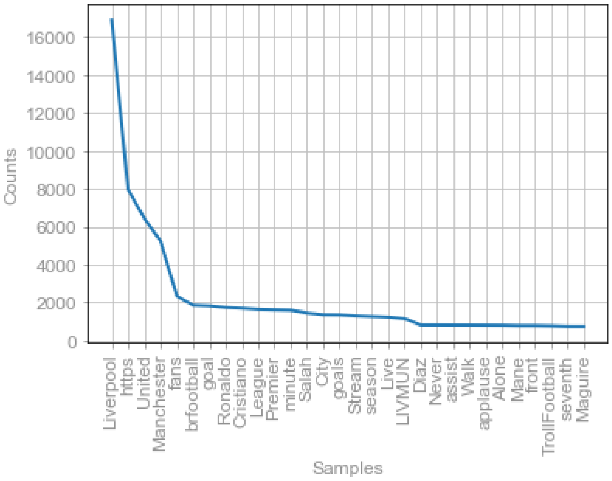


Figure 5. Frequency count of words in tweets

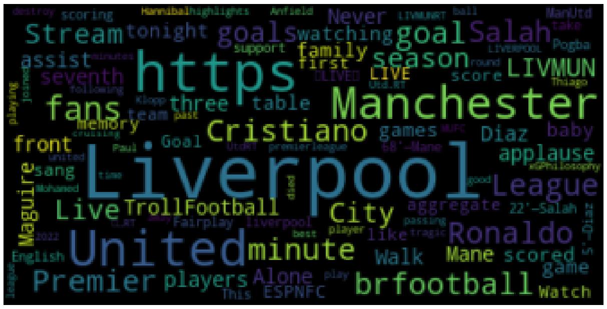


Figure 6. Wordcloud of words in tweets

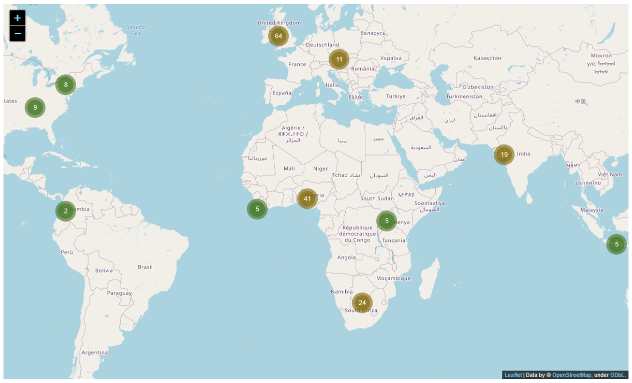


Figure 7. Location Of Tweets Worldwide

retweet count but a very low like. This may imply a targetted campaign to reach a wider audience. Since, it's common for tweets to have more likes than retweets.

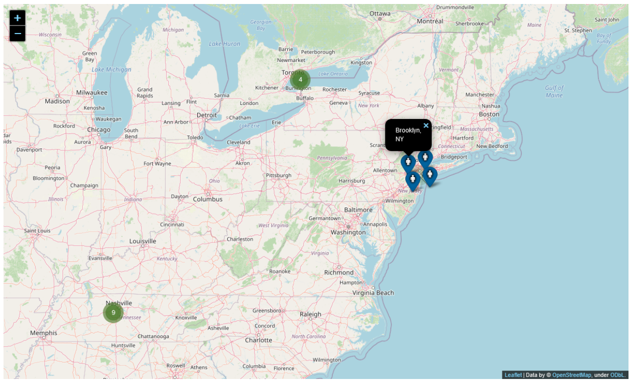


Figure 8. Location Of Tweets America

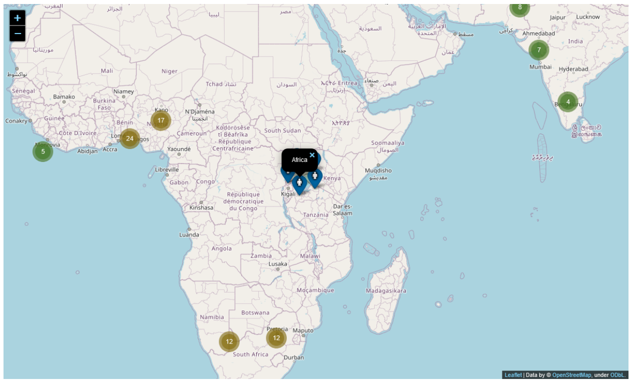


Figure 9. Location Of Tweets Africa

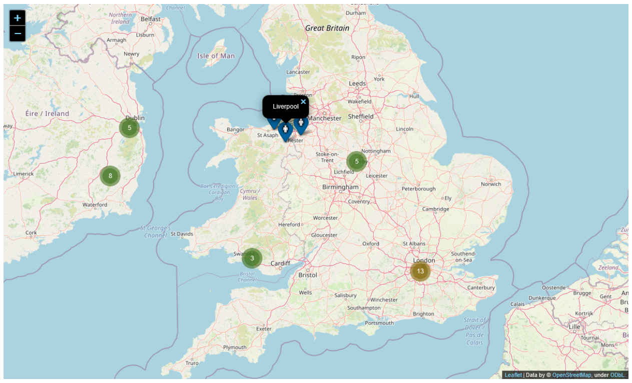


Figure 10. Location Of Tweets, United Kingdom

2.7. Followers & Like

Fig 12, shows relation, b/w Followers and Likes. It's worth mentioned that the x-axis units are measure in scientific notation. Therefore, generally it can be seen that tweets that have large following also have many likes on their tweets.

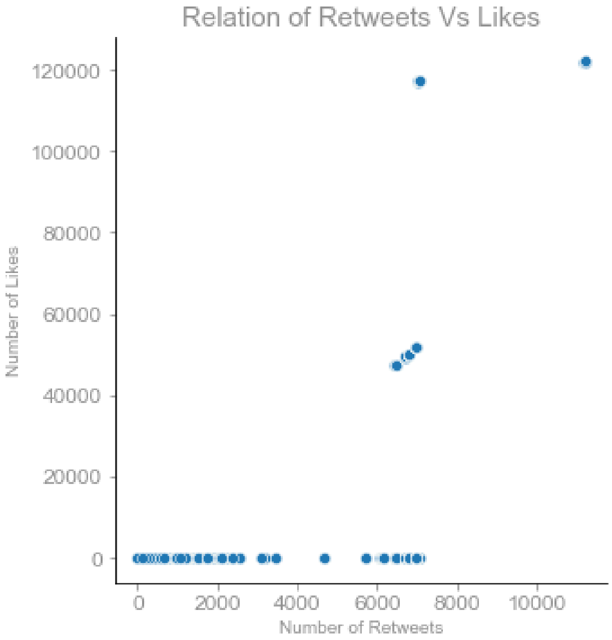


Figure 11. Relation of Retweets Likes

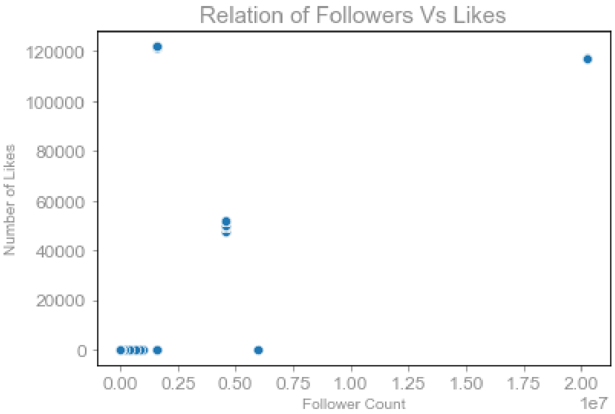


Figure 12. Relation of Followers Likes

3. Network & Community Detection

Community Detection is also an important aspect of Social Media Analysis. A social network of an individual is considered as its interaction & relationship with other users. The process of discovering the communities and cluster them together is called as community detection.

The Twitter data for community detection analysis is taken from (Leskovec & Krevl, 2014). The analysis is done in two steps.

- Network Analysis
- Community Detection Algorithms.

3.1. Network Analysis

Network analysis was done to check the number of nodes, edges, centrality measures of the Twitter data. The table represents the key findings of the Twitter data;

Table 1. Twitter Data Network Statistics

Nodes	226.
Edges	3634
Average Connectivity	5,932
Average Degree Centrality	0.14293
Shortest Path 1st to last	3 edges

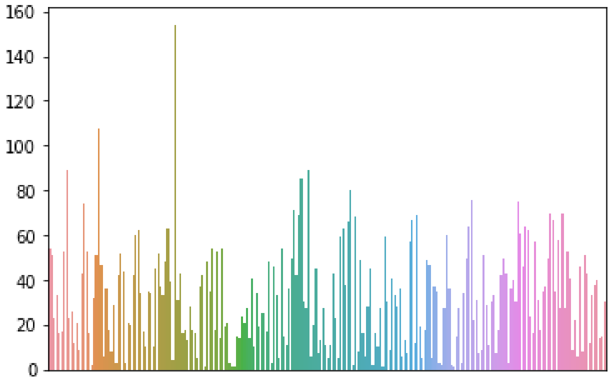


Figure 13. Degree of individual nodes

Table 1 shows the statistics of the data. The data contains 226 nodes and has, 3634 edges or connections. The average connectivity of the entire network is 5,932 which shows average number of connections within the network.

Degree centrality represents the direct link a node has with other nodes. The average degree of centrality of the network is 0.14293. Also, the amount of edges to traverse from the first to last node is 3. Fig 13 represents the nodes that have the highest connectivity to other nodes.

3.1.1. Betweenness Centrality

Betweenness Centrality helps to measure the mediation role of a node. Meaning, how many times a node is traversed when moving between different nodes. The Betweenness Centrality of the Twitter data is represented below;

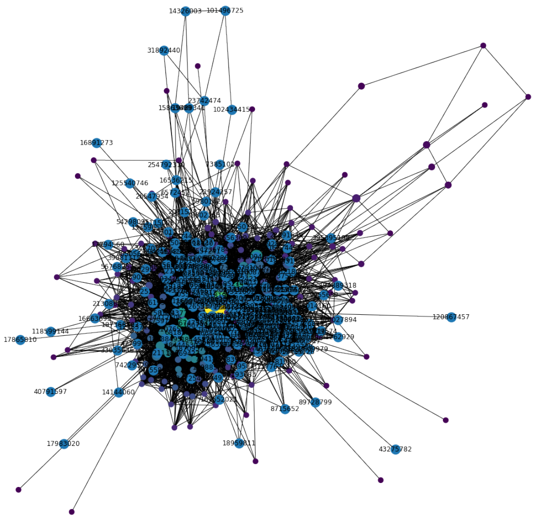


Figure 14. Betweenness Centrality of twitter data nodes

In fig 14, larger nodes show high betweenness, whereas the darker nodes show high degree of centrality. Average betweenness of the whole network is 0.01355. Nodes 16038438, 7861312, 1183041, 7081402, 20661527 had the high betweenness centrality.

3.1.2. Eigenvector Centrality

Eigenvector Centrality measures influence of a node on a network. It gives value based on how a node is connected to other important nodes. The Eigenvector Centrality of the Twitter data is represented below;

In fig 15, larger nodes show high Eigen value, whereas the darker nodes show high degree of centrality. Average eigenvector centrality of the whole network is 0.05258. Nodes 16038438, 7861312, 21158690, 130897520, 20747847 had the high eigenvector centrality.

3.2. Community Detection Analysis

Community Detection analysis is used to cluster or group the similar nodes together. The clustering is done based on the similarity and how closely the nodes are connected to each other. For community detection, ”CDLib” library was used to cluster the nodes. ”CDLib” has different algorithms:

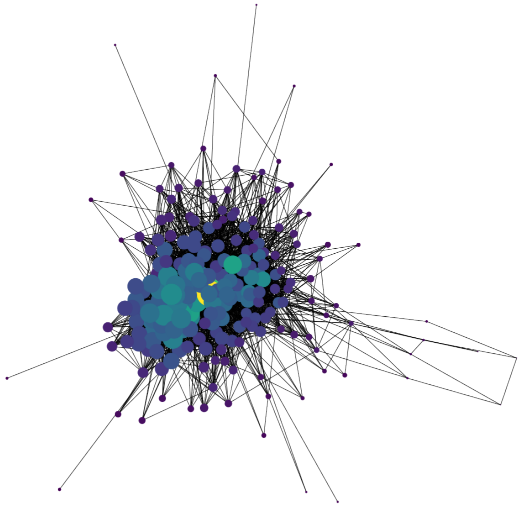


Figure 15. Eigenvector Centrality of twitter data nodes

Leiden, Louvain, Walktrap, Infomap and Label Propagation. These algorithms use different distance metrics to cluster the nodes.

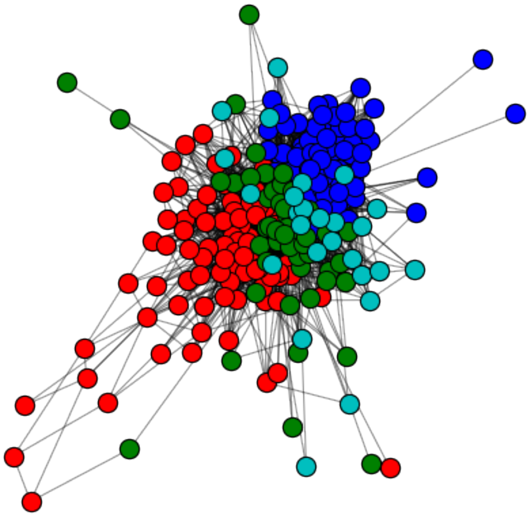


Figure 16. Louvain Community Detection of Twitter data

Both fig 17, 16 highlight 4 communities based on the distance metric used by the respective algorithms. However, it's worth noting that the result obtained from these algorithms are stochastic and multiple reruns may produce different communities

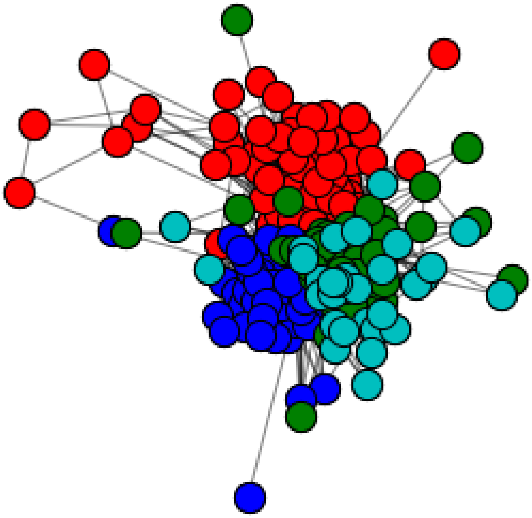


Figure 17. Leiden Community Detection of Twitter data

Features	Leiden	Louvain
Number Of Communities	4	4
Average Embeddedness of 4 Communities	0.751, 0.755, 0.386, 0.426	0.750, 0.755, 0.495, 0.388
Average Distance in Communities	2.03, 1.55, 1.85, 1.99	2.033, 1.568, 1.859, 1.913

Table 3.2 represents the statistical results of the Louvain Leiden algorithm on Twitter data. There were 4 communities detected by both algorithms. Average within cluster distance that was used to group nodes together was 2.03, 1.55, 1.85, 1.99 for Leiden Algorithm. Within cluster distance for Louvain Algorithm for 4 clusters was 2.033, 1.568, 1.859, 1.913.

Embeddedness describes how close the neighbours of a vertex are to being a clique, and this can be measured by calculating the local clustering coefficient. Average embeddedness of the clusters represent within similarity of nodes.

The fig 18 represents how similar different algorithms cluster the data together compared to each other. From the

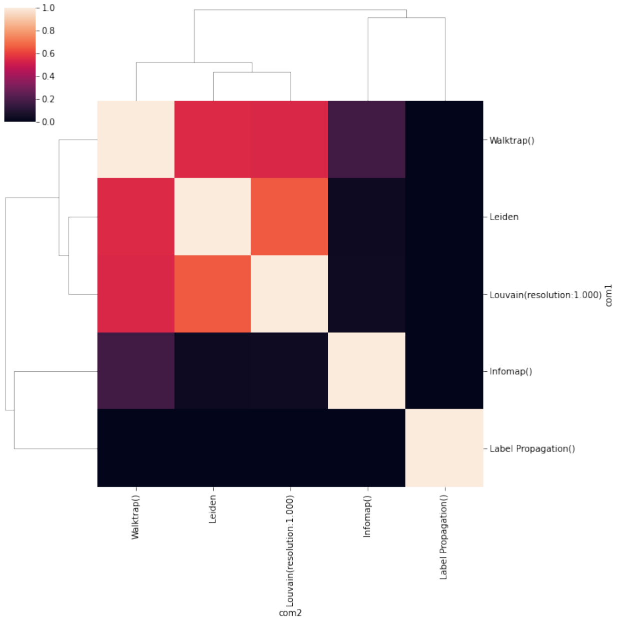


Figure 18. Similarity Matrix of all algorithms

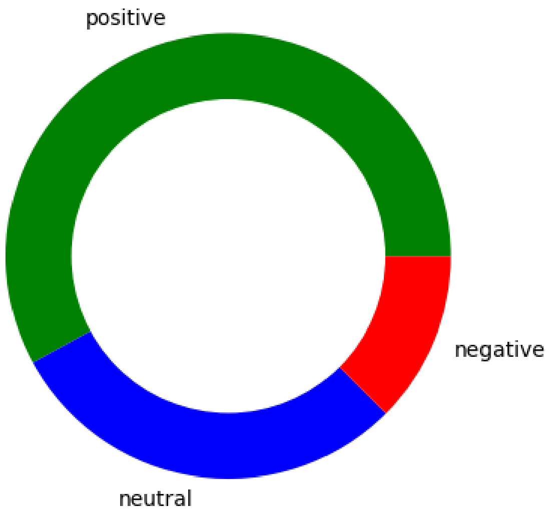


Figure 19. Sentiments percentage in tweets

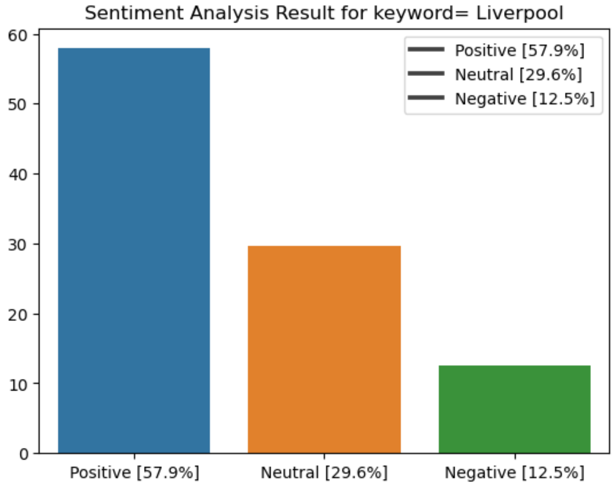


Figure 20. Sentiment Analysis of tweets

Table 2. Number & Percentage of Sentiment in Tweets

Sentiment	Total	Perc.
Positive	11354	57.93
Negative	5798	29.58
Neutral	2448	12.49

matrix Walktrap algorithm has most affinity with Leiden & Louvain algorithm.

4. Sentiment Analysis Of Tweets

Sentiment analysis on the content of the tweets was performed by using ”nltk” library. ”SentimentAnalyzer” class was used to classify tweets into ”pos”, ”neg” and ”neu” class. ”Pre-Processing” was performed on the tweet content. Regular expression was used to replace ”RT @” and any ”http links” with an empty space. Stopwords removal & lemmatization was used in preprocessing. Below are the results of the sentiments relating to”Liverpool” trend:

Table 2, fig 19, 20, 21 highlight that approx. 58% of the tweets had positive sentiment. Approx. 30% of tweets had negative words in it. Meanwhile, a small percentage of 12.49% tweets were neutral.

4.1. Word cloud & Frequency of Different Sentiments

Tweets that had positive, negative & neutral sentiment scores were extracted into separate lists. Each list was used to generate a wordcloud and frequency distribution plot to find the type of words used in each sentiment.

4.1.1. Tweets With Positive Sentiment

Fig 23, 22 highlight the words ”Liverpool”, ”enjoy”, ”watching”, ”goal”, ”walk”,”family”, ”live”, ”sang”, ”watch” are the words that made the sentiments of the tweets positive.

4.1.2. Tweets With Negative Sentiment

Fig 24, 25 highlight the words ”nothing”, ”useless”, ”died”, ”fighting”, ”defeat” are the words that made the senti-

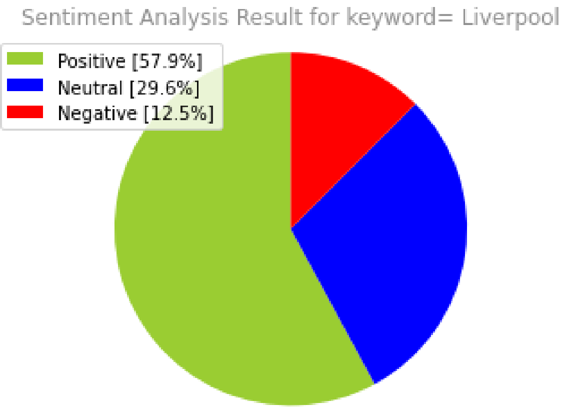


Figure 21. Percentage of Sentiments in tweets

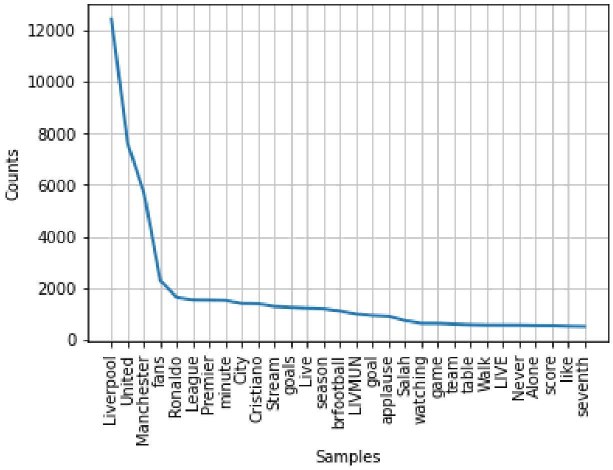


Figure 22. Word Frequency Distribution Positive Sentiment tweets

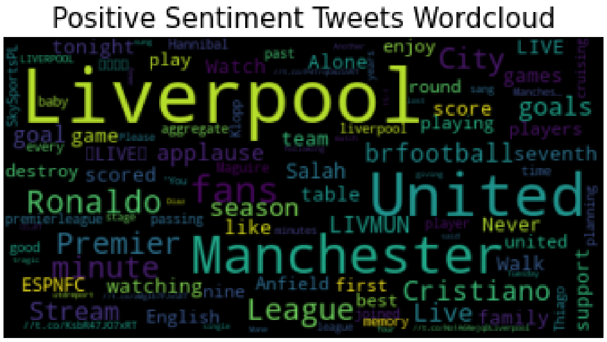


Figure 23. Word Cloud of Positive Sentiment tweets

ments of the tweets positive. To give the context, Cristiano Ronaldo's newborn baby died before the match (theguardian).

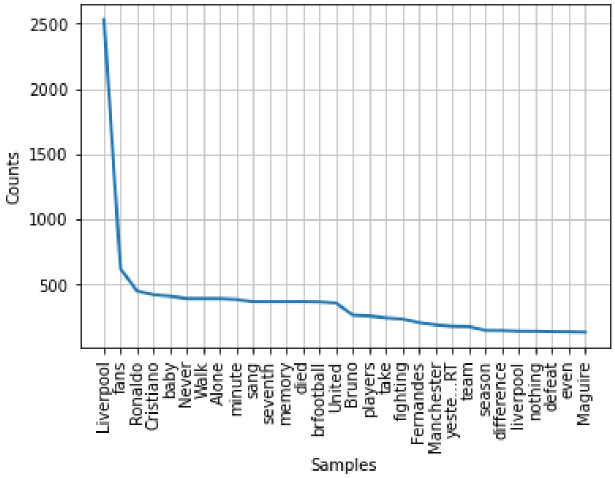


Figure 24. Word Frequency Distribution Negative Sentiment tweets

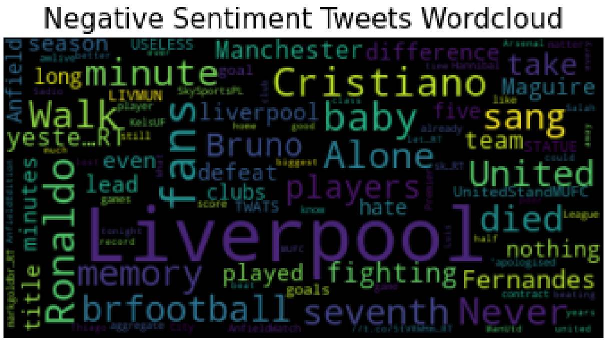


Figure 25. Wordcloud Of Negative Sentiment tweets

4.1.3. Tweets With Neutral Sentiment

Fig 26, 27 highlight the word frequency distribution & wordcloud of tweets categorized as neutral.

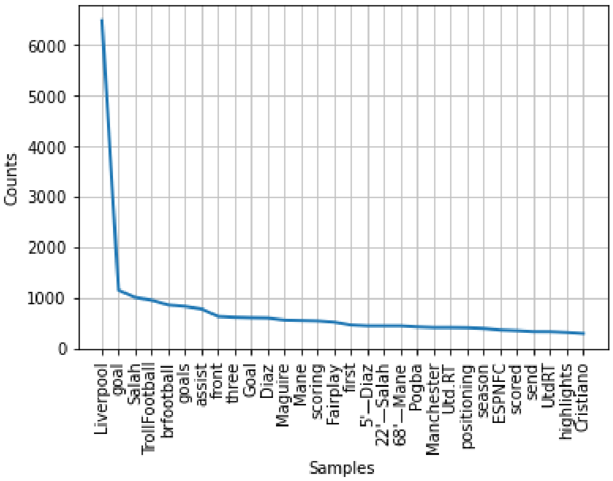


Figure 26. Word Frequency Distribution Neutral Sentiment tweets

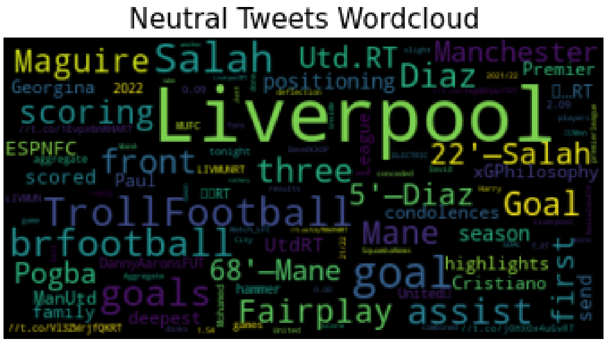


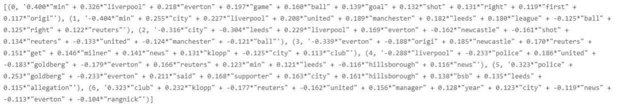
Figure 27. Wordcloud Of Neutral Sentiment tweets

5. Newspaper Articles

Topic modelling is a type of statistical technique in Machine Learning and NLP, that is used to discover ”Topic” in a document. For topic modelling, News API was used to get the articles related to ”Liverpool”. A total of 10 articles were extracted. The URL of each article was then used to dynamically web scrape the article content from their original website.

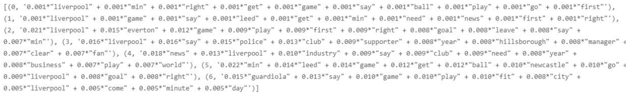
Both Latent Semantic Analysis (LSA) & Latent Dirichlet allocation (LDA) algorithm from ”Genism” package were used. LSA assumes that similar meaning words will appear in similar piece of texts. LDA assumes that each topic is a mixture of an underlying set of words and each document is a mixture set of topic probabilities.

Libraries used in this case were ”Genism”, ”NLTK”, ”Spacy”, ”pyLDAvis”. The LSA & LDA were used on all 10 articles text, preprocessing involved Lemmatization & stopwords removal and total of 7 topics were selected. Fig 28 & 29 show LSA and LDA lists that include tuple of 7 proposed topics. Both algorithms come up with similar topics likes ”Liverpool”, ”City”, ”Everton”. This can be expected as Liverpool & Man City are currently competing for English Premier League.



```
[(0, '0.400*"min" + 0.326*"liverpool" + 0.218*"everton" + 0.197*"game" + 0.160*"ball" + 0.139*"goal" + 0.132*"shot" + 0.131*"right" + 0.119*"first" +
0.117*"origi"'), (1, '-0.404*"min" + 0.255*"city" + 0.227*"liverpool" + 0.208*"united" + 0.189*"manchester" + 0.182*"leeds" + 0.180*"league" + -0.125*"ball" +
0.125*"right" + 0.122*"reuters"'), (2, '-0.316*"city" + -0.304*"leeds" + 0.229*"liverpool" + 0.169*"everton" + -0.162*"newcastle" + -0.161*"shot" +
0.134*"reuters" + -0.133*"united" + -0.124*"manchester" + -0.121*"ball"'), (3, '-0.339*"everton" + -0.188*"origi" + 0.185*"newcastle" + 0.170*"reuters" +
0.151*"get" + 0.146*"milner" + 0.141*"news" + 0.131*"klopp" + -0.125*"city" + 0.113*"club"'), (4, '-0.288*"liverpool" + -0.233*"police" + 0.186*"united" +
-0.183*"goldberg" + -0.179*"everton" + 0.166*"reuters" + 0.123*"min" + 0.121*"leeds" + -0.116*"hillsborough" + 0.116*"news"'), (5, '0.323*"police" +
0.253*"goldberg" + -0.233*"everton" + 0.211*"said" + 0.168*"supporter" + 0.163*"city" + 0.161*"hillsborough" + 0.138*"bsb" + 0.135*"leeds" +
0.115*"allegation"'), (6, '0.323*"club" + 0.232*"klopp" + -0.177*"reuters" + -0.162*"united" + 0.156*"manager" + 0.128*"year" + 0.123*"city" + -0.119*"news" +
-0.113*"everton" + -0.104*"rangnick"')]
```

Figure 28. LSA Topic Models



```
[(0, '0.001*"liverpool" + 0.001*"min" + 0.001*"right" + 0.001*"get" + 0.001*"game" + 0.001*"say" + 0.001*"ball" + 0.001*"play" + 0.001*"go" + 0.001*"first"'),
(1, '0.001*"liverpool" + 0.001*"game" + 0.001*"say" + 0.001*"leed" + 0.001*"get" + 0.001*"min" + 0.001*"need" + 0.001*"news" + 0.001*"first" + 0.001*"right"'),
(2, '0.021*"liverpool" + 0.015*"everton" + 0.012*"game" + 0.009*"play" + 0.009*"first" + 0.009*"right" + 0.008*"goal" + 0.008*"leave" + 0.008*"say" +
0.007*"min"'), (3, '0.016*"liverpool" + 0.016*"say" + 0.015*"police" + 0.013*"club" + 0.009*"supporter" + 0.008*"year" + 0.008*"hillsborough" + 0.008*"manager" +
0.007*"clear" + 0.007*"fan"'), (4, '0.018*"news" + 0.013*"liverpool" + 0.010*"industry" + 0.009*"say" + 0.009*"club" + 0.009*"need" + 0.008*"year" +
0.008*"business" + 0.007*"play" + 0.007*"world"'), (5, '0.022*"min" + 0.014*"leed" + 0.014*"game" + 0.012*"get" + 0.012*"ball" + 0.010*"newcastle" + 0.010*"go" +
0.009*"liverpool" + 0.008*"goal" + 0.008*"right"'), (6, '0.015*"guardiola" + 0.013*"say" + 0.010*"game" + 0.010*"play" + 0.010*"fit" + 0.008*"city" +
0.005*"liverpool" + 0.005*"come" + 0.005*"minute" + 0.005*"day"')]
```

Figure 29. LDA Topic Models

Fig 30, 31, 32, 33, 34 show the visualization of the topics suggested by LDA algorithm for articles 1, 2, 3, 4, 5.

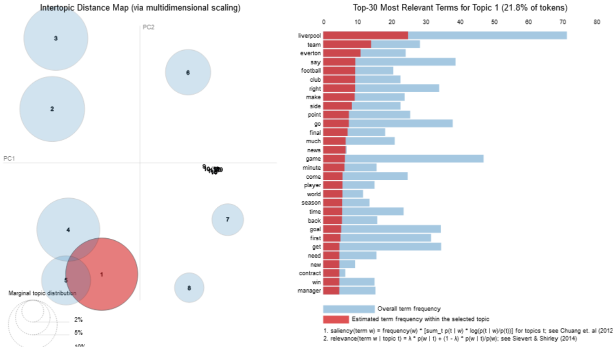


Figure 30. LDA Topic Models For Article 1

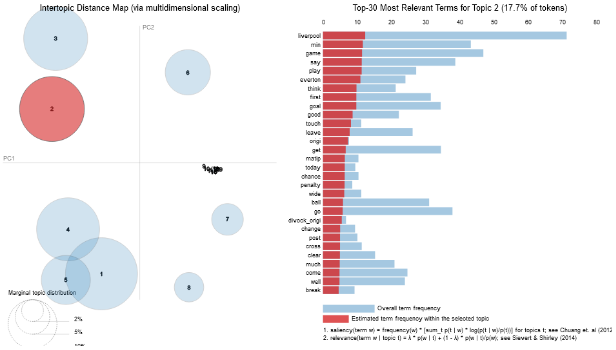


Figure 31. LDA Topic Models For Article 2

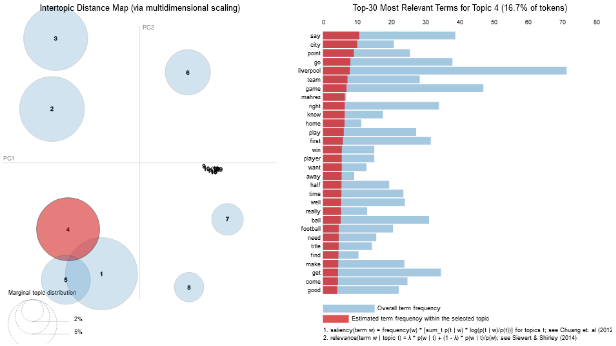


Figure 32. LDA Topic Models For Article 3

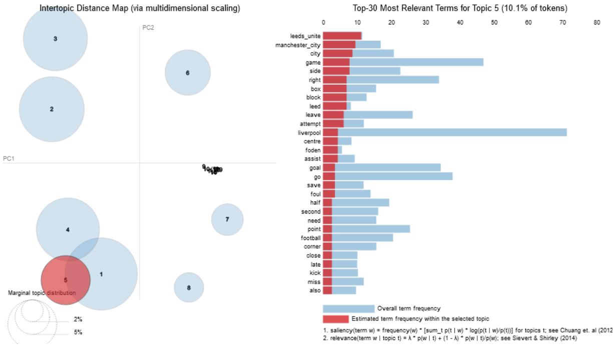


Figure 33. LDA Topic Models For Article 4

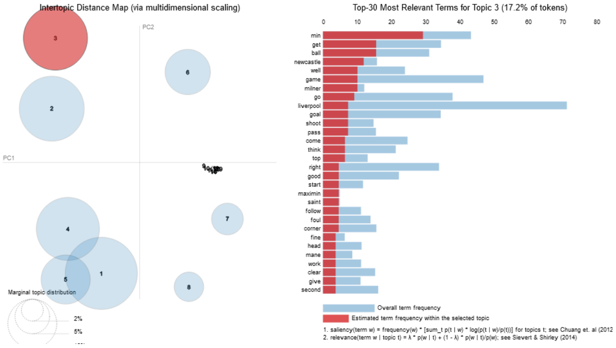


Figure 34. LDA Topic Models For Article 5

5.1. Time Series of Word Count in Articles

Fig 35 show the word count in articles escalate on dates 19th, 24th & 30th of April. Those were also the dates when Liverpool played matches against Everton, Newcastle and Man City. This suggests articles on match days tend to have high word count as people are more willing to spend their time. All the links to articles are available at appendix A.

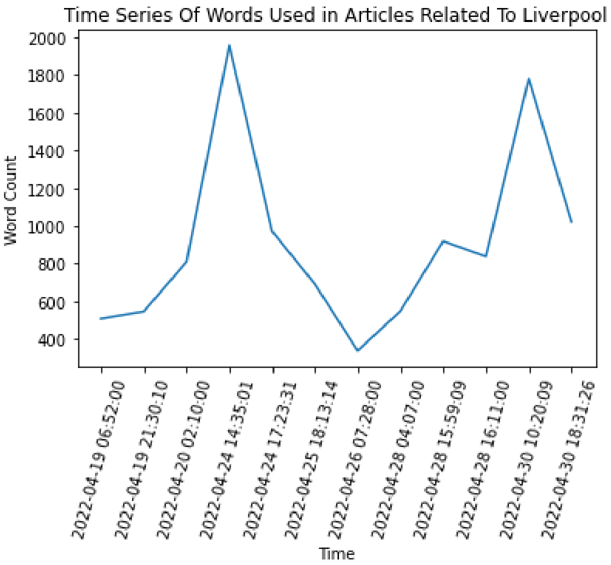


Figure 35. Time Series Of Word Counts Used in Articles For Liverpool Keyword

5.2. WordCloud Of Articles

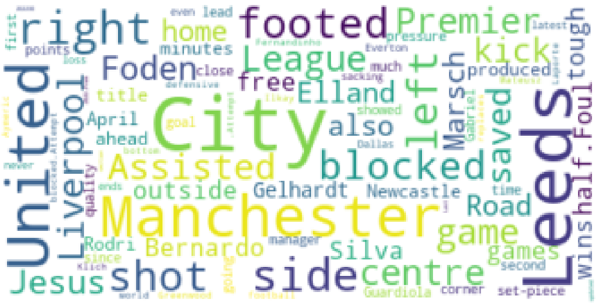


Figure 36. Wordcloud Of Article 1

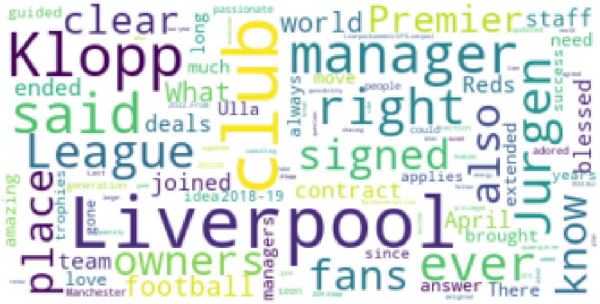


Figure 37. Wordcloud Of Article 2

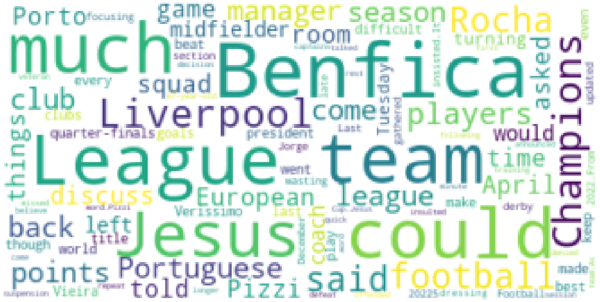


Figure 38. Wordcloud Of Article 3

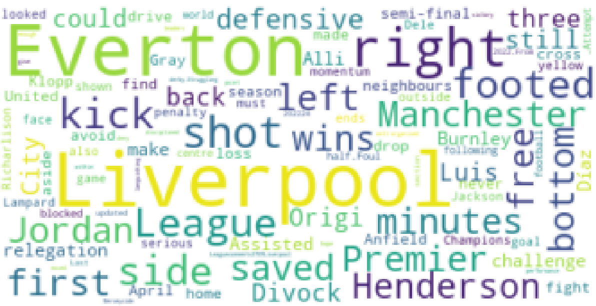


Figure 39. Wordcloud Of Article 4

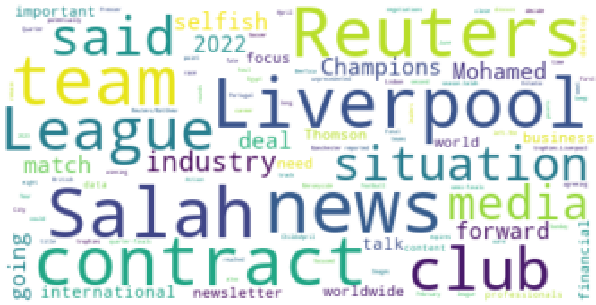


Figure 40. Wordcloud Of Article 5

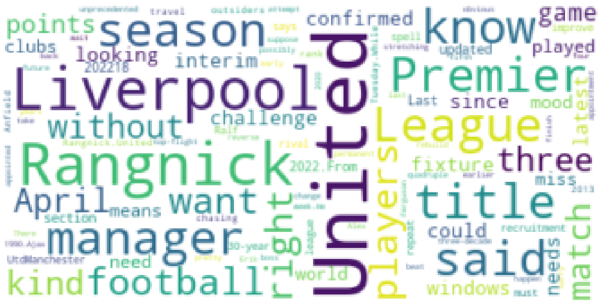


Figure 41. Wordcloud Of Article 6

5.3. Frequency Distribution Of words in Articles

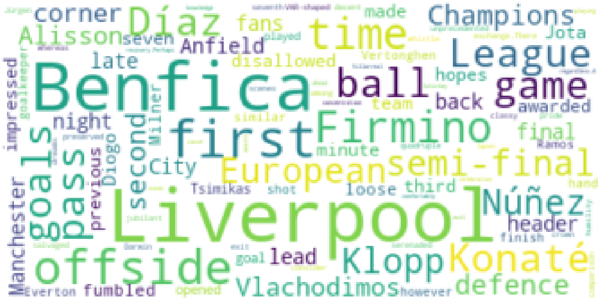


Figure 42. Wordcloud Of Article 7

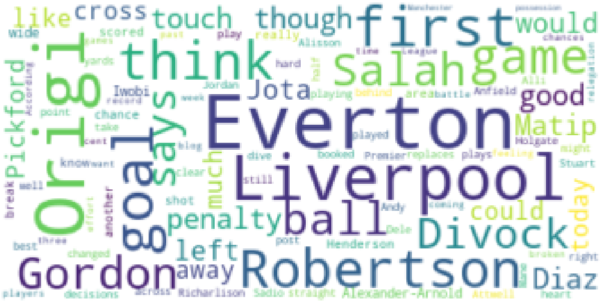


Figure 45. Wordcloud Of Article 10

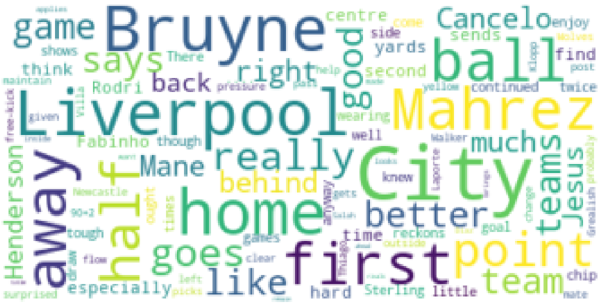


Figure 43. Wordcloud Of Article 8

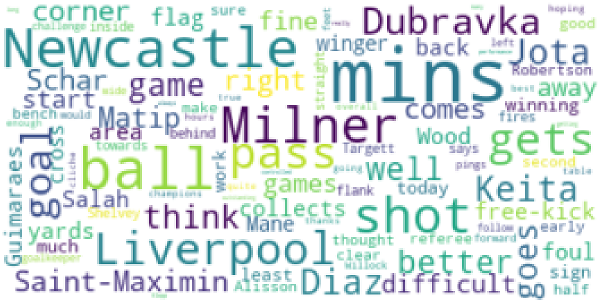


Figure 44. Wordcloud Of Article 9

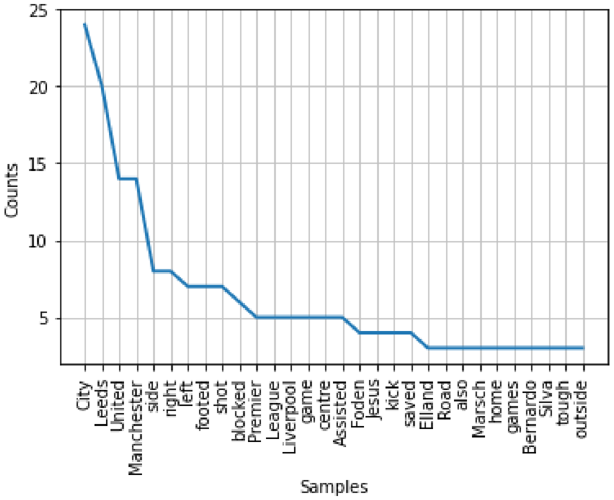


Figure 46. FreqDist Of Article 1

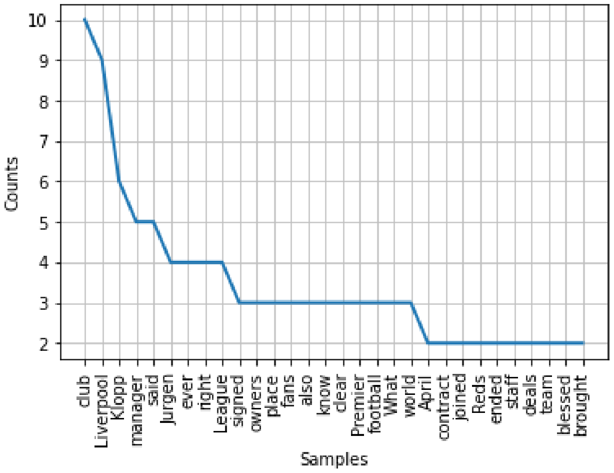


Figure 47. FreqDist Of Article 2

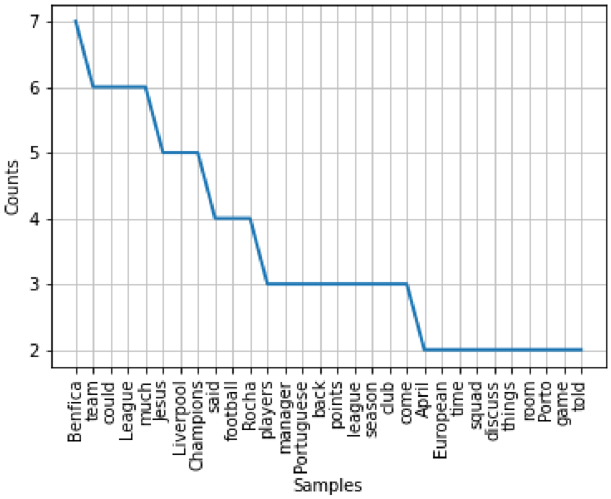


Figure 48. FreqDist Of Article 3

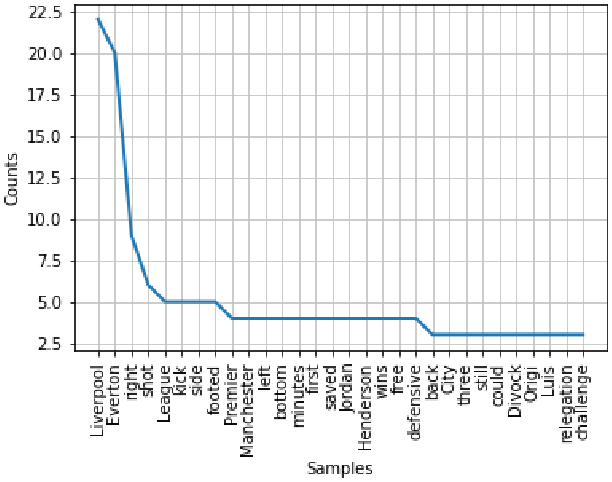


Figure 49. FreqDist Of Article 4

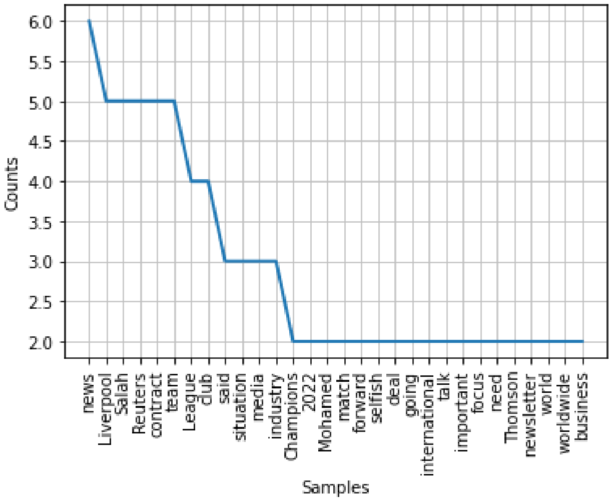


Figure 50. FreqDist Of Article 5

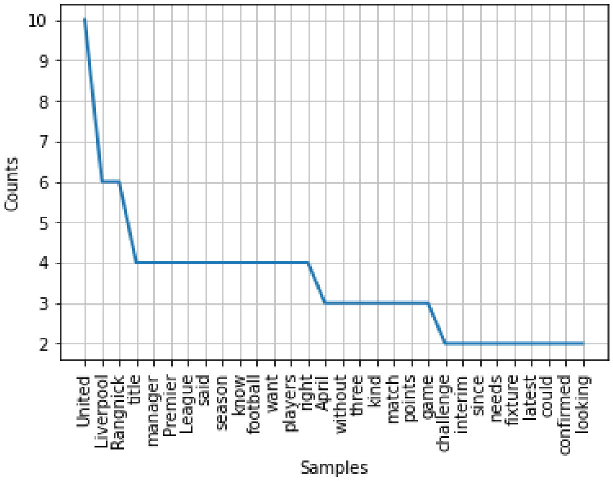


Figure 51. FreqDist Of Article 6

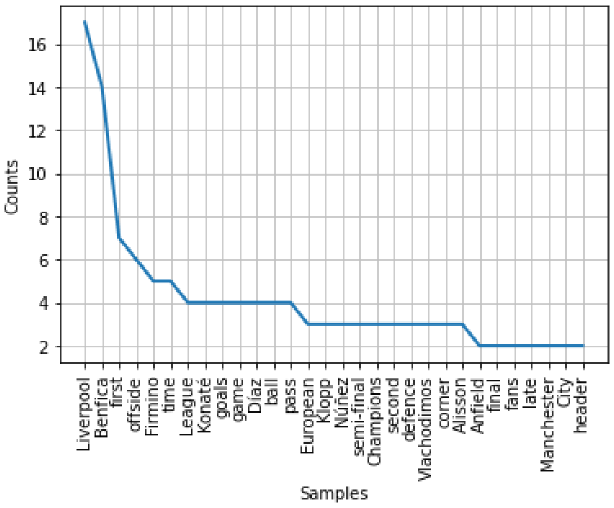


Figure 52. FreqDist Of Article 7

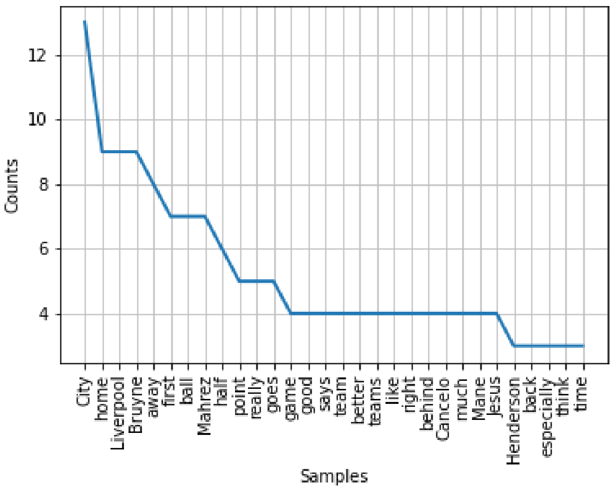


Figure 53. FreqDist Of Article 8

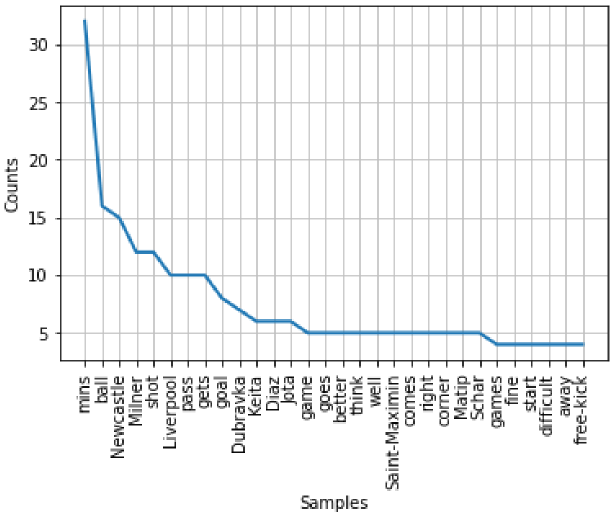


Figure 54. FreqDist Of Article 9

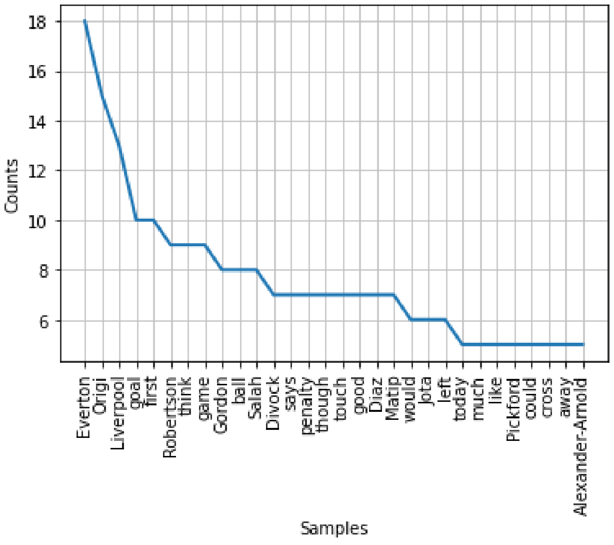


Figure 55. FreqDist Of Article 10

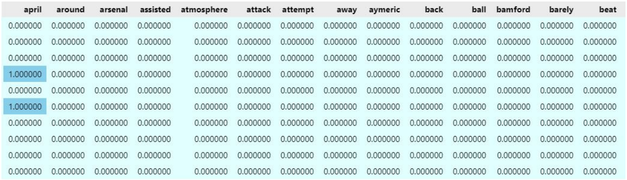
5.4. Bag of Words (BOW) & Term Frequency Inverse document frequency TF IDF

Table 3 show the Article's title & wordcount. The first article A was used for document summarization, BOW & TF IDF. The whole article contained 35 sentence tokens and, 1020 words. Both BOW & TF IDF analysis give a matrix of 457 rows x 296 columns, representing 457 sentences and 296 different words left after preprocessing.

Table 3. Article Titles & their Word Count

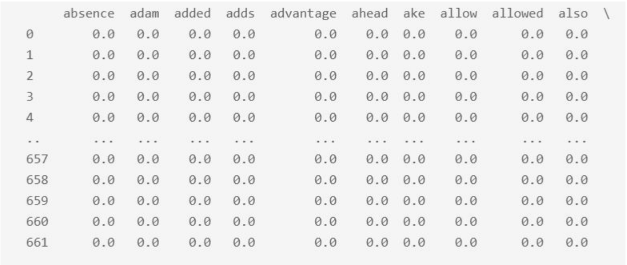
Article Title	Word Count
Leeds United 0-4 Manchester City...	1020
Jurgen Klopp: Liverpool...	918
Benfica v Liverpool...	886
Liverpool 2-0 Everton...	971
Salah says he cannot be...	528
Manchester United will... - Ralf Rangnick	661
Liverpool to face ...	871
Manchester City v Liverpool...	1367
Newcastle United v Liverpool...	1779
Liverpool v Everton...	1957

Fig 57, 56 show the TF IDF of the first article. Fig 59, 58 show the BOW of A article. In both BOW, TF IDF, it can be assumed that not every word is present in most sentences. However, the word ”April” is in most sentences that make sense, as most matches and articles are written in the month of April.



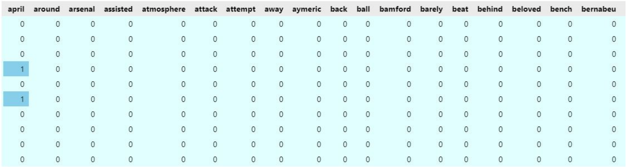
april	around	arsenal	assisted	atmosphere	attack	attempt	away	aymeric	back	ball	bamford	barely	beat
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 56. Snapshot of TF IDF of Article 1



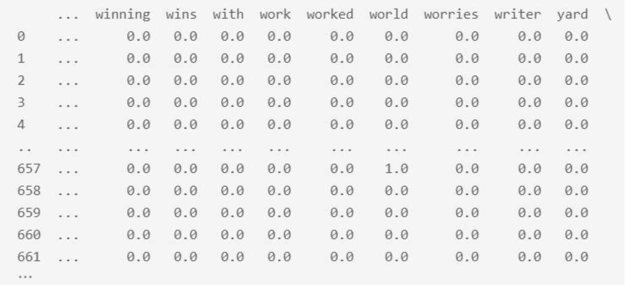
```
     absence  adam  added  adds  advantage  ahead  ake  allow  allowed  also  \
0        0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
1        0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
2        0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
3        0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
4        0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
..       ...   ...    ...   ...        ...    ...  ...    ...      ...   ...
657      0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
658      0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
659      0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
660      0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
661      0.0   0.0    0.0   0.0        0.0    0.0  0.0    0.0      0.0   0.0
```

Figure 57. TF IDF of Article 1



april	around	arsenal	assisted	atmosphere	attack	attempt	away	aymeric	back	ball	bamford	barely	beat	behind	beloved	bench	bernabeu
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 58. Snapshot of BOW of Article 1



```
     ...  winning  wins  with  work  worked  world  worries  writer  yard  \
0    ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
1    ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
2    ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
3    ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
4    ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
..   ...      ...   ...   ...   ...     ...    ...      ...     ...   ...
657  ...      0.0   0.0   0.0   0.0     0.0    1.0      0.0     0.0   0.0
658  ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
659  ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
660  ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
661  ...      0.0   0.0   0.0   0.0     0.0    0.0      0.0     0.0   0.0
...
```

Figure 59. BOW of Article 1

5.5. Topic Summarization

An article summary was generated by using the sentence tokenization and weighted score of word tokens and their relative frequency in the sentence. The words were first tokenized and weighted based on their document frequency. The first article A was used for document summarization. The whole article contained 35 sentence tokens and, 1020 words. and 5% of the document was selected to generate the following summary:

”For the late game, manager Pep Guardiola had to consider not only the challenges posed by the traditional fearsome atmosphere inside Elland Road and a Leeds United side desperate for the points, but also this coming Wednesday's Champions League semi-final second leg against Real Madrid.Guardiola left Kevin de Bruyne, Bernardo Silva and Riyad Mahrez on the bench to keep them fresh for the test of protecting a 4-3 lead in Spain.The strategy worked as City came through in relative comfort in what was always going to be a tough physical examination.”

The summary above perfectly describes the in-match & post match events that occurred during the Leeds vs Man City match.

6. Justification & Limitations of the used methods

6.1. Justification of choices

All the techniques used in this analysis involved following preprocessing steps: stopwords removal, regex to remove symbols, special characters and text lemmatization to mitigate the effect of any noise on the models.

Additionally, the report tries to give a holistic overview by analysing the latest trends on Twitter, what client platforms people use. What are the dominant sentiments in the Tweets. Wordcloud & Frequency distribution plots were used to emphasize the important words used. Donut plot & tables were used to check the percentages of sentiment and number of tweets in each sentiment.

A comparison of community detection algorithms was drawn on Twitter data. Key nodes, Centrality comparison & betweenness were measured. The report then used 10 different articles to perform topic modelling techniques. Time Series analysis, Wordcloud analysis. The report used sentence tokenization and word weightage to generate a summary of the first article.

6.2. Limitations

The report is limited as Twitter trends are dynamic. Also, the effort is further hampered with limit on requests allowed by Twitter. Furthermore, not all tweets include location information, as these are dependent on the user permissions. Lastly, the quality of results for relationship comparison are dependent on the types of tweets & articles fetched from the API.

7. Conclusion

The report gives a holistic overview of Twitter trends, Sentiment Analysis, Newspaper Topic modelling. Intensive data cleaning & pre-processing were applied to get the quality representation. The report accomplishes all the tasks required for analysis. All the representation revealed the insights and were more importantly true. All the code used is available at appendix B.

8. Appendix

A. Links of articles used in Newspaper analysis

1- https://www.bbc.co.uk/sport/football/61198621,2022-04-30T18:31:26Z
2- https://www.bbc.co.uk/sport/football/61262331,2022-04-28T15:59:09Z
3- https://www.bbc.co.uk/sport/football/60980975
4- https://www.bbc.co.uk/sport/football/61131364
5- https://www.reuters.com/lifestyle/sports/salah-says-he-cannot-be-selfish-discuss-his-liverpool-contract-situation-2022-04-09/
6- https://www.bbc.co.uk/sport/football/61142369
7- https://amp.theguardian.com/football/2022/apr/13/liverpool-benfica-champions-league-quarter-final-second-leg-match-report
8- https://amp.theguardian.com/football/2022/apr/13/liverpool-benfica-champions-league-quarter-final-second-leg-match-report
9- https://www.theguardian.com/football/live/2022/apr/30/newcastle-v-liverpool-premier-league-live-score-updates
10- https://www.theguardian.com/football/live/2022/apr/24/liverpool-v-everton-premier-league-live

B. Code Available at Github

1- https://github.com/xahram/websmassessmentnotebook

References

Ali, A. From Amazon to Zoom: What Happens in an Internet Minute In 2021, url = ”https://www.visualcapitalist.com/from-amazon-to-zoom-what-happens-in-an-internet-minute-in-2021/”.

Daniel, D. Tweeting the revolution. 2013. URL http://www.culturaldiplomacy.org/pdf/case-studies/daniel_domingues_-_tweeting_the_revolution.pdf.

Dean, B. How Many People Use Twitter in 2022 [New Twitter Stats, url = ”https://backlinko.com/twitter-users”.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014. URL https://snap.stanford.edu/data/twitter.tar.gz.

theguardian. Cristiano Ronaldo and Georgina Rodríguez announce death of baby son, url = ”https://www.theguardian.com/football/2022/apr/18/cristiano-ronaldo-and-georgina-rodriguez-announce-death-of-baby-son”.