CMP7202 Web Social Media Analytics and Visualisation

Dr Ogerta Elezaj

Joshua Fernandes
22169738

Word count- 1978

Contents

Introduction

This report is split into 5 parts. The first is the reddit post analysis which will look at trending subreddits like cybersecurity. Data will be pulled from the API, analysed, and then visualised. The second part will feature graph analysis and will look at Degree analysis, betweenness centrality and eigenvector centrality. It will also include community detection. The third part will be the sentiment analysis for comments made on the subreddit for cybersecurity. The final part will look at the news API analysis and will consist of topic modelling and Article summarisation.

Reddit post Analysis

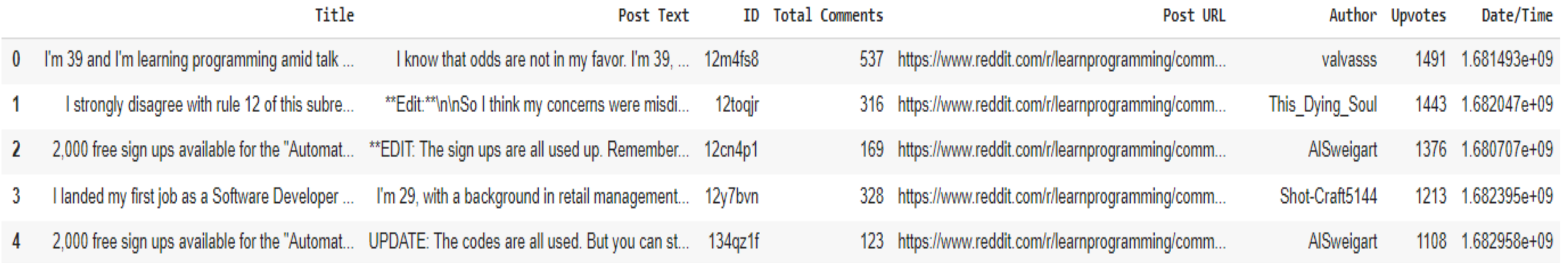
Reddit is a social network with a forum-style discussion structure where content is curated by the public using voting and awards (Stafford, 2016). The site name is a play on the words ‘I read it’. Users create posts in topic-based communities known as subreddits. Each thread on a subreddit has an original poster which redditors refer to as OP. A subreddit serves as a mini community like Facebook groups except for the fact that they are public.

Trending Subreddits related to Cybersecurity.

Using the Reddit API data has been pulled from the top 10 subreddits related to Cybersecurity as shown in table 1. The columns pulled from the Reddit API are shown in Figure 2. These posts were collated to make one data frame consisting of all those posts.

r/cybersecurity
r/informationtechnology
r/python
r/privacy
r/hacking
r/dataanalysis
r/askprogramming
r/technology
r/learnprogramming
r/learnpython

Table 1 List of subreddits to scrape



	Title	Post Text	ID	Total Comments	Post URL	Author	Upvotes	Date/Time
0	I'm 39 and I'm learning programming amid talk ...	I know that odds are not in my favor. I'm 39, ...	12m4fs8	537	https://www.reddit.com/r/learnprogramming/comm...	valvasss	1491	1.681493e+09
1	I strongly disagree with rule 12 of this subre...	**Edit:**\n\nSo I think my concerns were misdi...	12toqjr	316	https://www.reddit.com/r/learnprogramming/comm...	This_Dying_Soul	1443	1.682047e+09
2	2,000 free sign ups available for the "Automat...	**EDIT: The sign ups are all used up. Remember...	12cn4p1	169	https://www.reddit.com/r/learnprogramming/comm...	AlSweigart	1376	1.680707e+09
3	I landed my first job as a Software Developer ...	I'm 29, with a background in retail management...	12y7bvn	328	https://www.reddit.com/r/learnprogramming/comm...	Shot-Craft5144	1213	1.682395e+09
4	2,000 free sign ups available for the "Automat...	UPDATE: The codes are all used. But you can st...	134qz1f	123	https://www.reddit.com/r/learnprogramming/comm...	AlSweigart	1108	1.682958e+09

Figure 1 List of columns in data frame

Word Cloud and Frequency Distribution for post Titles

Figure 3 shows a word cloud for the prevalent words in the Title column. The data has first been cleaned and the word Python has been removed as it is the name of one of the subreddits and this would have an impact on the wordcloud.Figure 4 shows the frequency distribution of the words. As is visible from the figure words like ‘data’,’code’ and ‘analyst’ were repeated frequently as one would expect in technology related subreddits. Other words that were prevalent were words like ‘use’, ‘learn’ and ‘job’ as a lot of subreddits focus on teaching how to code.

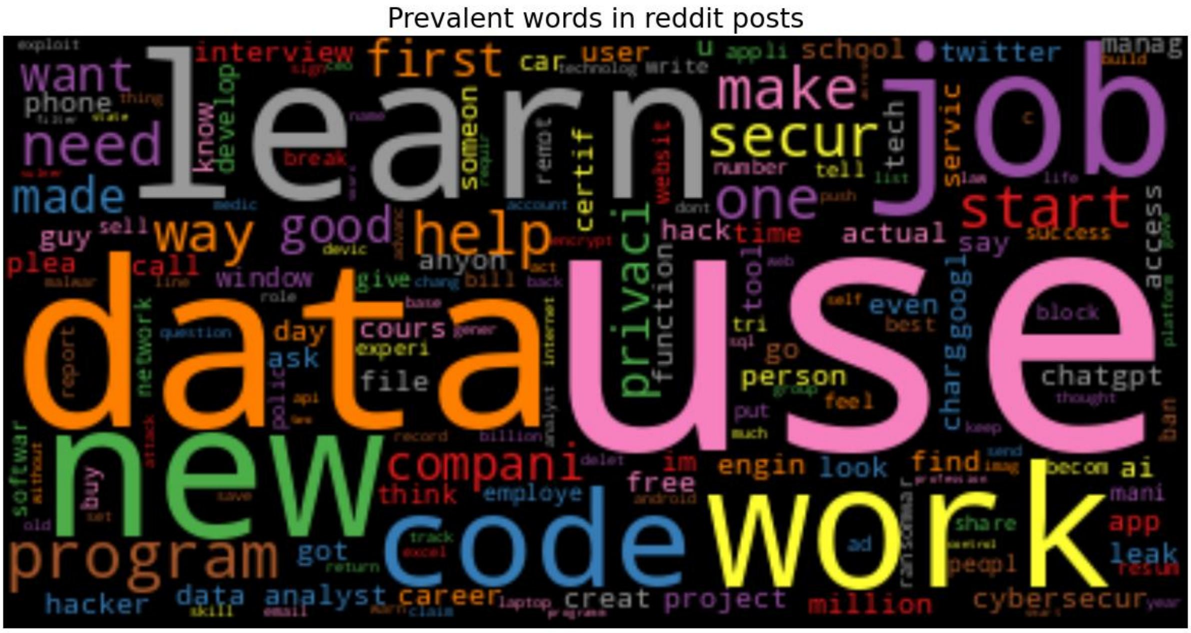


FIGURE 2 WORD CLOUD OF PREVALENT WORD IN THE TITLE OF REDDIT POSTS

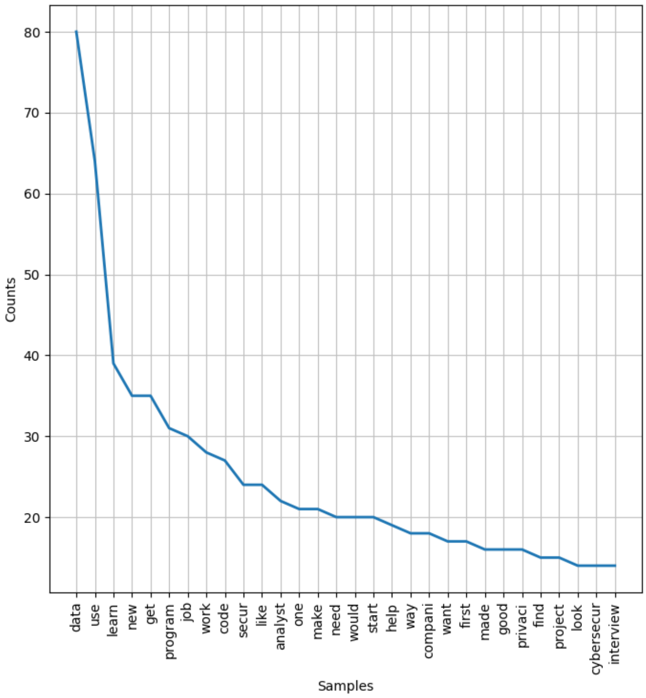


FIGURE 3 WORD FREQUENCY DISTRIBUTION FOR TITLE OF REDDIT POSTS

Missing Values

Figure 5 shows a heatmap of the missing values from the posts dataframe. As is visible from the figure only 2 columns have missing values Author and Post Text. Author has very few missing values and the cause of this is when the Author deletes their account. On the other hand, Post Text has many missing values caused by the fact that sometimes there is no body to the reddit posts as it may just be a link or image with a title.

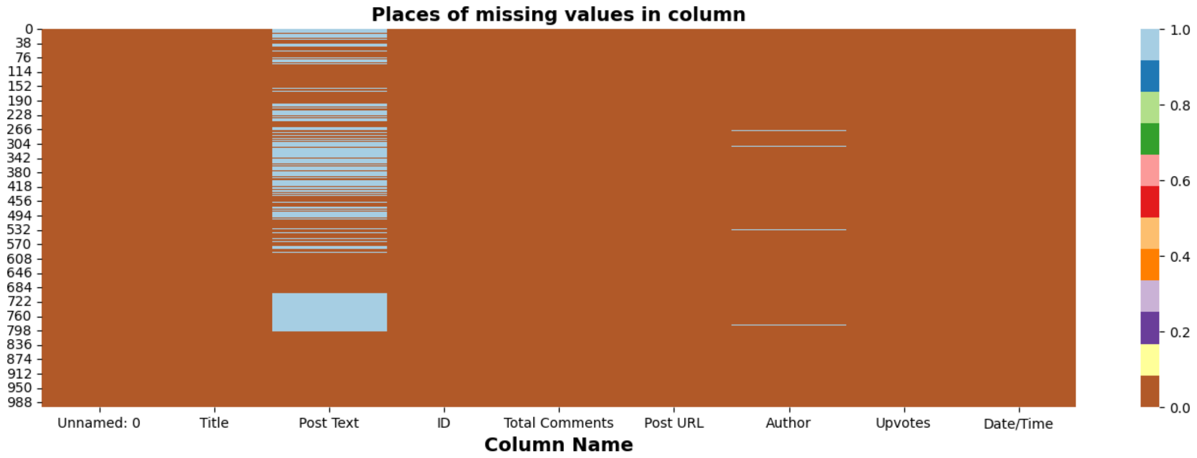


FIGURE 4 HEATMAP OF MISSING VALUES

Posts with the most upvotes

Figure 5 shows the 10 posts with the most upvotes. The data has been displayed in a horizontal bar chart since the label has a long name. Many of these articles show links of things happening in the news. This demonstrates how Reddit can also be used to catch up on news the same way the trending tab on twitter does.

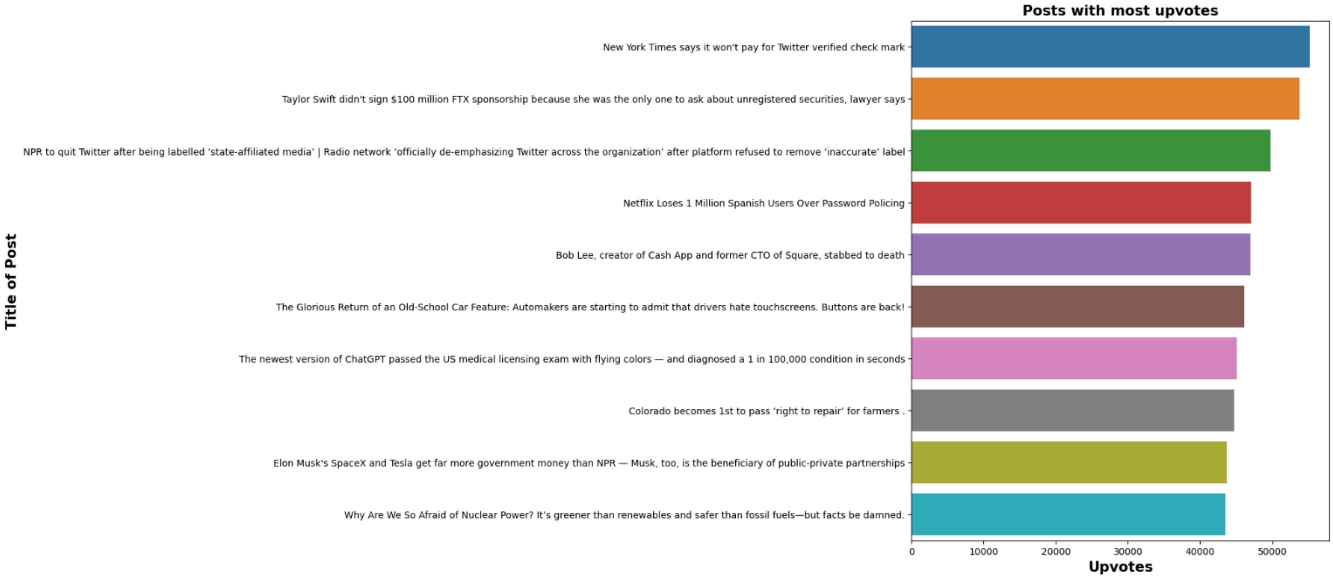


FIGURE 5 HORIZONTAL BAR CHART TO SHOW THE POSTS WITH HIGHEST UPVOTES

Authors with the most posts

Figure 6 shows the number of posts contributed by each author. These insights are important to identify active users in the community.This can give you an idea on the trustworthiness of a poster based on how much they post.

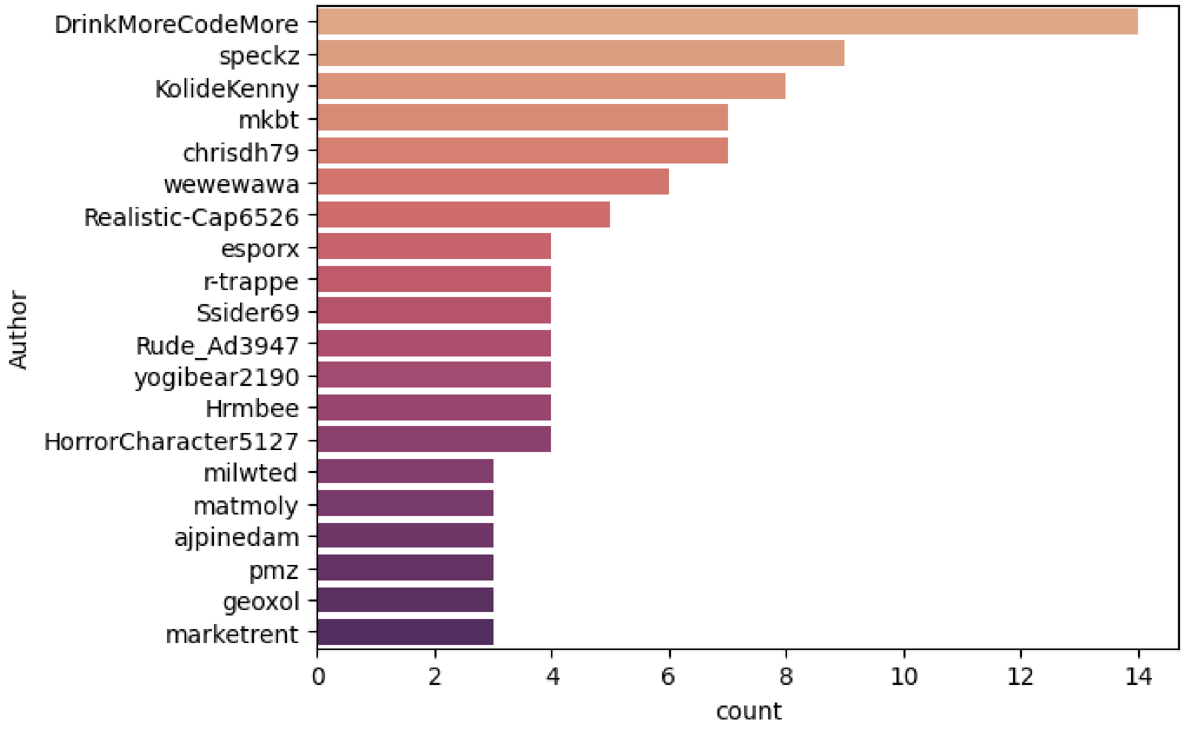


FIGURE 6 HORIZONTAL BAR CHART TO SHOW THE 20 MOST FREQUENT POSTERS

Reddit Post Date Time Analysis

Figure 7,8 and 9 show the number of by hour,date and day of the week respectively.Figure 8 shows that the most posts happened in the afternoon to evenings with 10 am being the outlier.Figure 9 shows that the most posts happened on the 12th April. Figure 10 shows that the most posting happens on a weekday rather than a weekend.

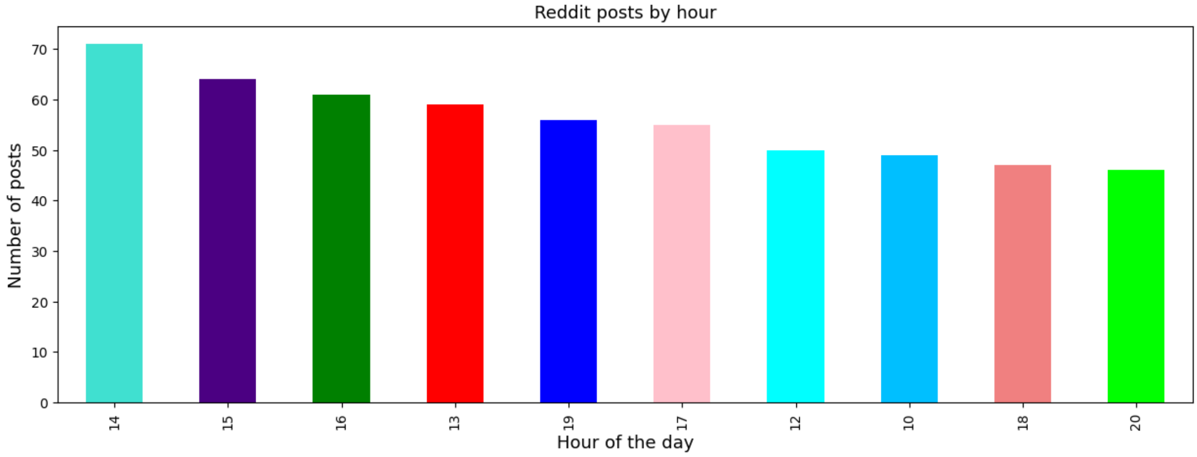


FIGURE 7 BAR CHART SHOWING THE 10 MOST ACTIVE POSTING HOURS

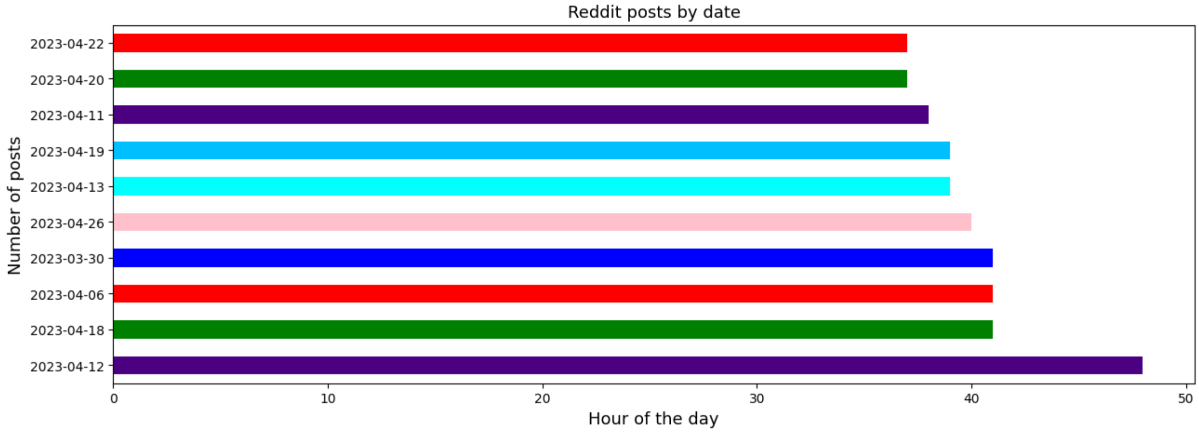


FIGURE 8 HORIZONTAL BAR CHART THAT SHOWS THE 10 DAYS WITH THE MOST POSTS IN A MONTH

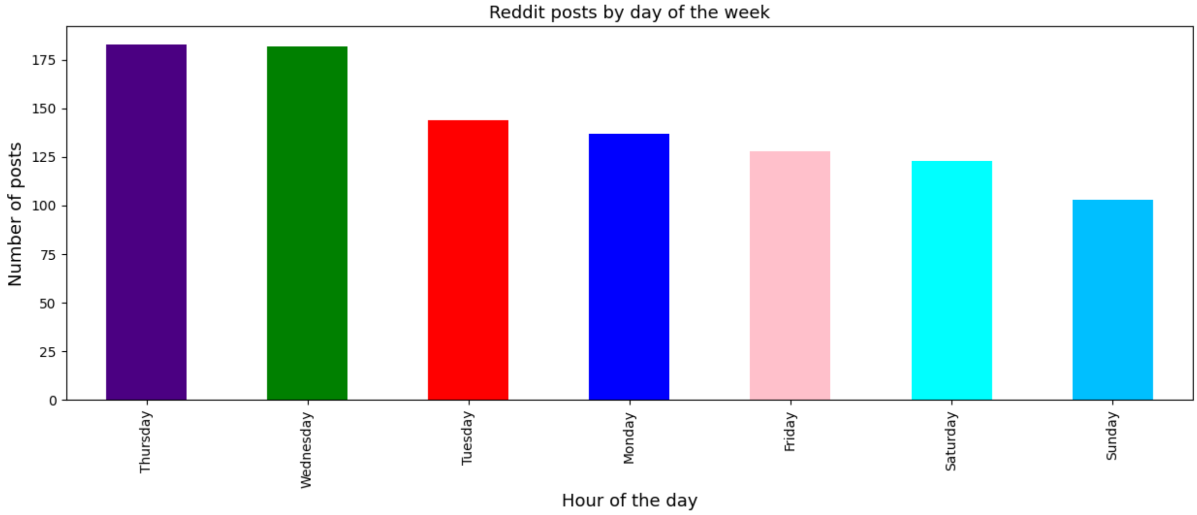


FIGURE 9 BAR CHART SHOWING THE DAYS OF THE WEEK WITH THE MOST POSTS

Graph Analysis

The graph used for this analysis is a dataset for Twitch streamers that stream in Portuguese.The nodes itself are the streamers themselves.The links between them are their mutual friendships.There are 1912 nodes and 31,299 edges.The analysis has been done using Gephi.

Degree Analysis

Figure shows graph visualisation based on the number of degrees. The 10 nodes with the highest degrees are shown in table 3.This shows the importance of a node based on how many edges are going in and out. The sum of the in degree and out degree is the total degree value shown. .Figure 14 shows the visualisation for the graph where the greater the degree the larger the circle is for the node.

Betweenness Centrality

Betweenness centrality captures how much a given node is in between other nodes.This is done by calculating by counting the number of shortest paths that go through the target node(Perez & Germon, 2016).Table 4 shows the nodes with the highest betweenness centrality whereas figure shows the distribution of the betweenness centrality data. As is visible from the distribution there are many counts of betweenness centrality that are low. This means that the graph isn't very well connected each other. Figure 14 shows the visualisation for the graph where the greater the betweenness centrality value the larger the circle is for the node.

Eigenvector Centrality

Eigenvector Centrality can be defined as the amount of influence a node has on the network(Golbeck, 2013). A fundamental concept of this is that if a node is connected to nodes with a higher score their eigenvector centrality will be higher(shaw, 2019). This means that if a node has a high degree but it's connections are with other nodes with low scores then the eigenvector centrality will be low. Table 5 shows the 10 nodes with the highest Eigenvector centrality. Node 1758 has an eigenvector centrality of 1.0 which is the highest possible value. This means that it is the dominant eigenvalue which can be defined as the eigenvalue that is greater than all the other eigenvalues. The distribution for the Eigenvector is clustered towards the beginning showing us that there are multiple eigenvalues with a low score. This means that there are multiple nodes that have very little influence to the network. Figure 15 shows the visualisation for the graph where the greater the value for eigenvector centrality is, the larger the circle is for the node.

Nodes	1912
Edges	31299
Average Degree	16.37
Minimum Degree	1
Maximum Degree	767

TABLE 2 DESCRIPTIVE ANALYTICS FOR GRAPH DATA

Node	Degree
127	767
1476	598
290	590
1297	587
467	582
1660	475
67	454
1320	416
1758	394
1259	385

TABLE 3 TOP 10 DEGREES

Node	Betweenness Centrality
1297	0.020505
467	0.018287
1476	0.018285
1660	0.015568
290	0.012848
127	0.010339
1259	0.009172
1014	0.007873
1311	0.007706
471	0.007407

TABLE 4 TOP 10 BETWEENNESS CENTRALITIES

Node	Eigenvector Centrality
1758	1.0
1320	0.906059
1593	0.831811
1821	0.76699
1476	0.737888
1787	0.710554
1739	0.706828
36	0.645908
1721	0.617972
1414	0.5306

TABLE 5 TOP 10 EIGENVECTOR CENTRALITIES

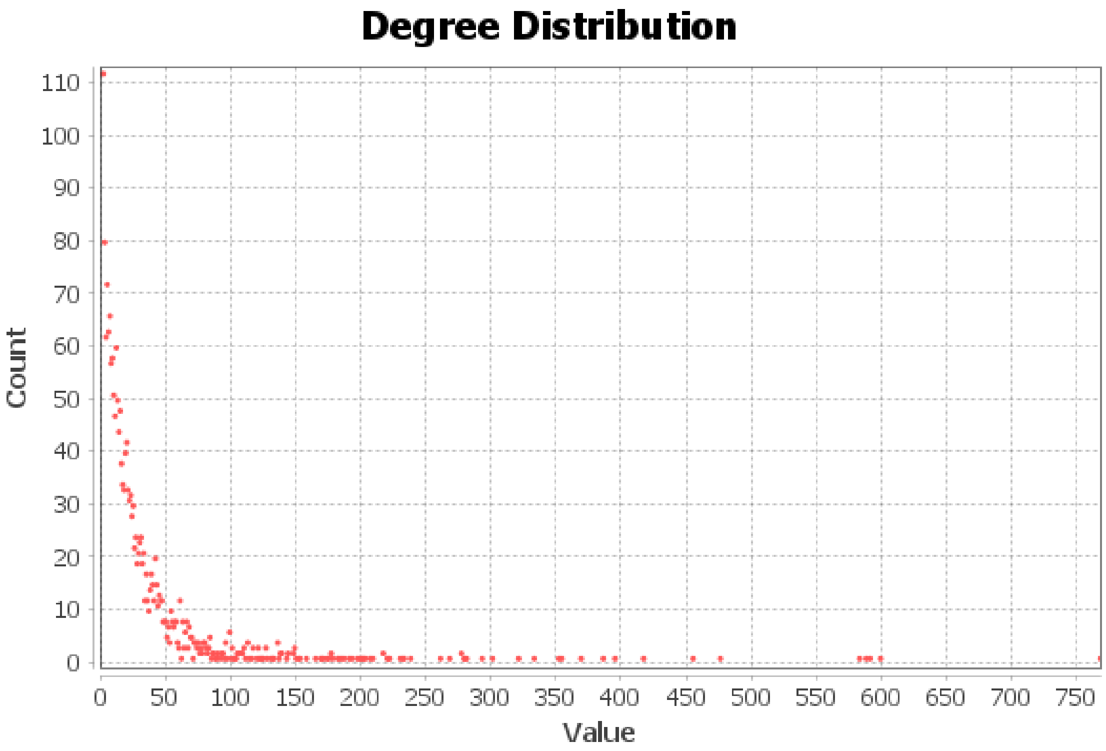


FIGURE 10 A GRAPH TO SHOW DEGREE DISTRIBUTION

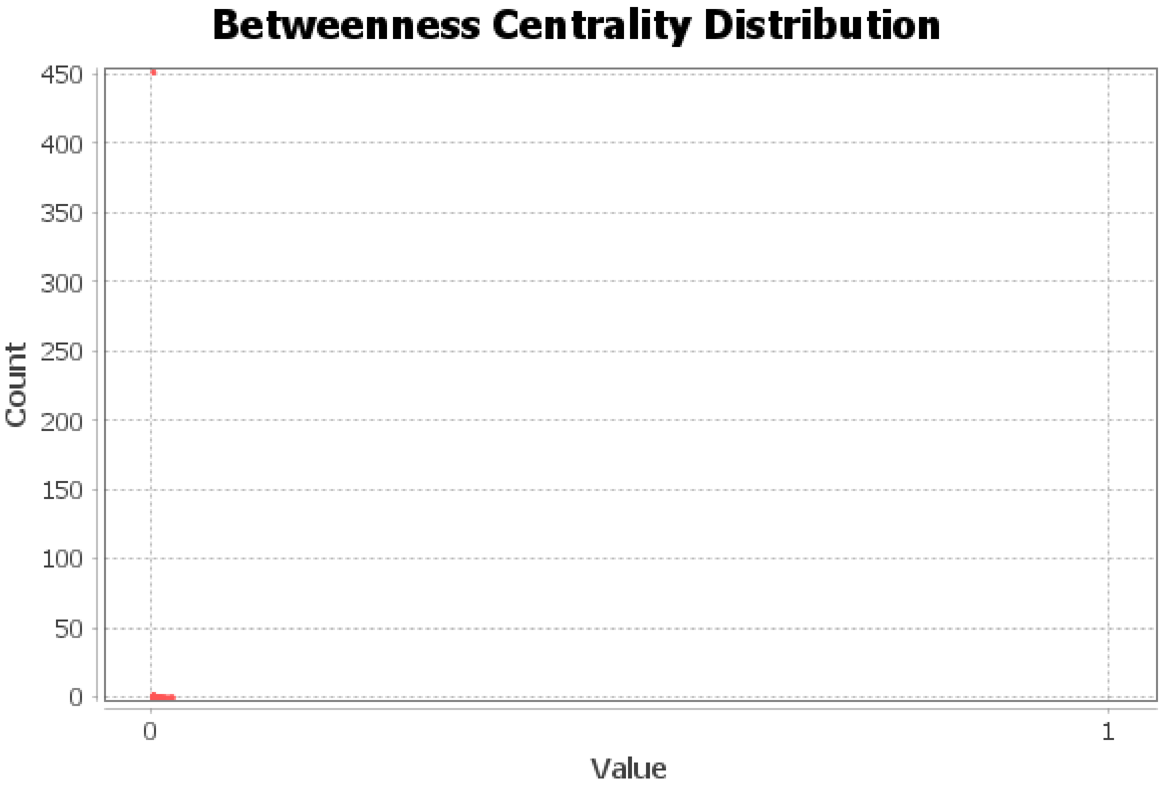


Figure 11 a graph to show betweenness centrality distribution

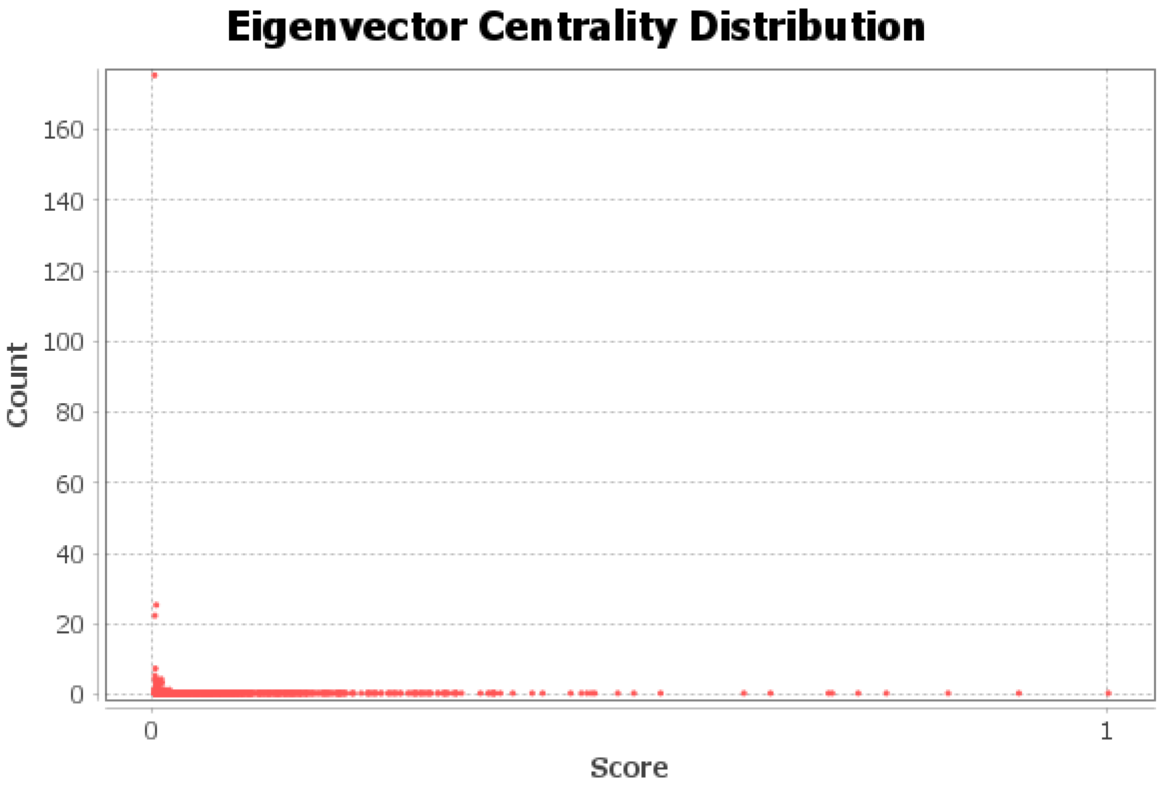


FIGURE 12 A GRAPH TO SHOW EIGENVECTOR CENTRALITY DISTRIBUTION

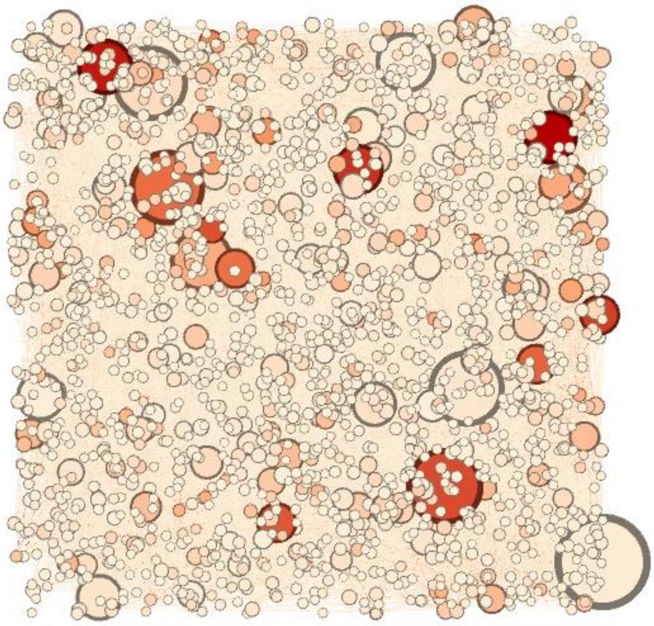


FIGURE 13 GRAPH VISUALISATION FOR DEGREE CENTRALITY

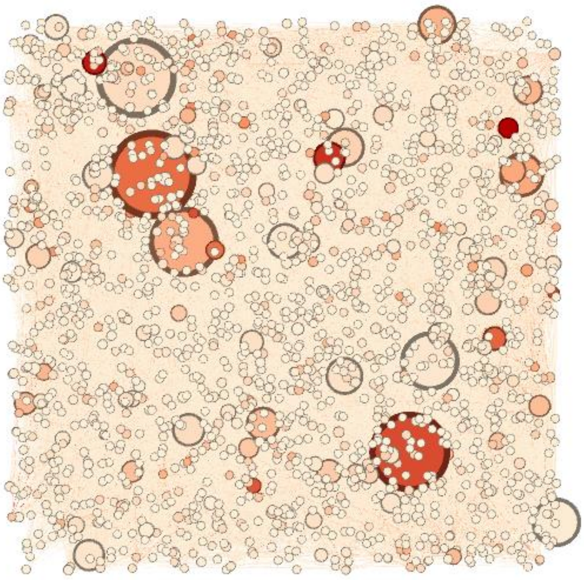


FIGURE 14 GRAPH VISUALISATION FOR BETWEENNESS CENTRALITY

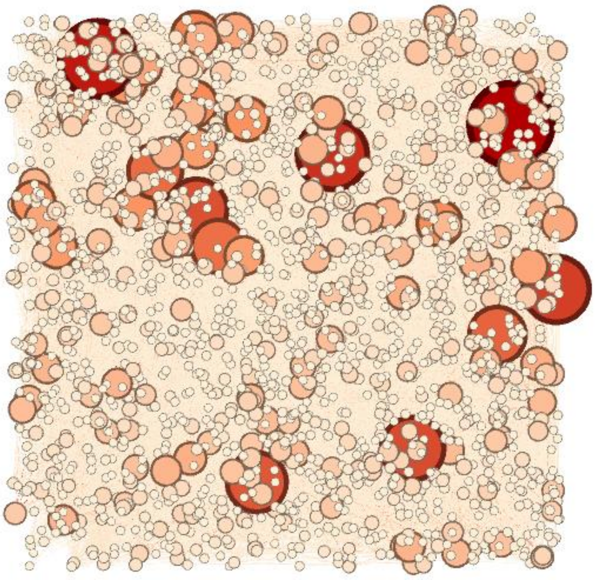


FIGURE 15 GRAPH VISUALISATION FOR EIGENVECTOR CENTRALITY

Community Detection

Gephi uses the Louvain algorithm to calculate community detection. Figure 16 shows a visualisation of the modular class which shows that there are 6 communities. Each node has a colour and the colour corresponds to what community it is. Table 6 shows the percentage of nodes in each community.The graph modularity score for this graph was 0.292. Graphs with a high modularity score means that they have many connections within their own communities.

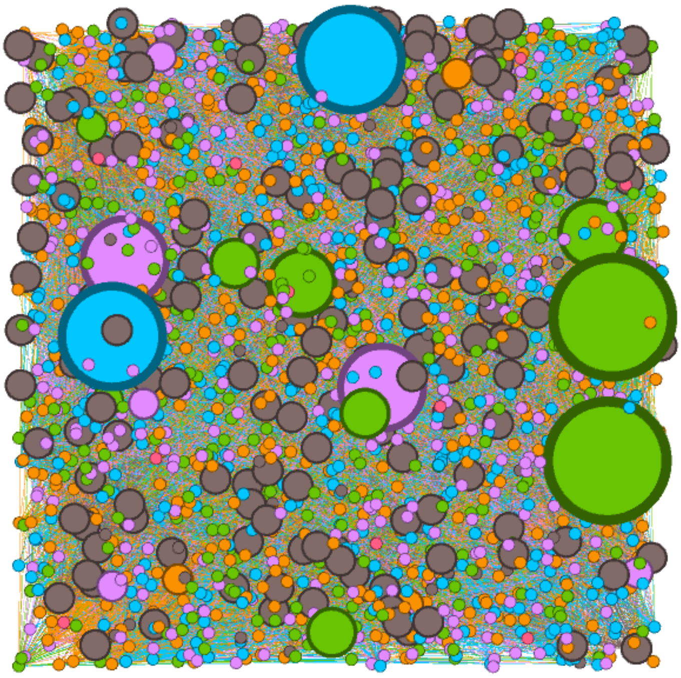


Figure 16 A graph to show community detection

Community	Percentage
Community 1(Orange)	27.14%
Community 2(Light Blue)	22.07%
Community 3(Red)	20.61%
Community 4(Green)	18.36%
Community 5(Dark Blue)	11.04%
Community 6(Pink)	0.78%

Table 6 Stats for each community within the graph

Sentiment Analysis for Reddit Comments

Comments collated for 11 different reddit posts in the cybersecurity subreddit. This was done by pulling them from the reddit API and transforming them into one data frame. Figure 17 displays a word cloud of the data. The prevalent words as shown in the word cloud are words like job, work and MFA which makes sense as most of the reddit posts have to do with either jobs in the cybersecurity sector, ways in which you get certified or educating redditors on their experiences.

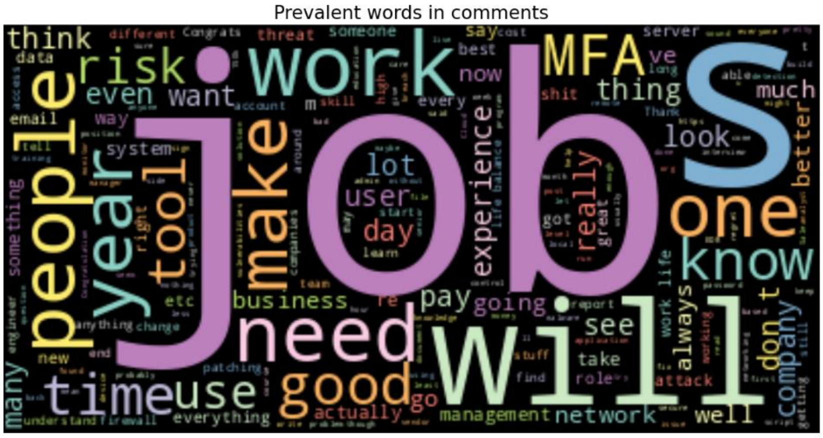


FIGURE 17 WORDCLOUD FOR WORDS IN REDDIT COMMENTS

Sentiment Analysis is done by using text blob. Text blob measures the intensity of the words and Is used to determine whether a word is positive, negative or neutral. In this case the comments were passed on to the text blob and using the text blob the polarity and subjectivity was calculated. If the polarity is greater than 0 the sentiment is positive. If the polarity is less than 0 it is negative and if the polarity is exactly 0 then it is neutral.

Figure 18 shows a bar chart of the distribution between positive, negative and neutral sentiment. As show by the bar chart it is evident that positive and neutral sentiment outweigh the negative sentiment, and this can be caused by the fact that the cybersecurity subreddit mainly focuses on teaching and giving other’s advice about cybersecurity.

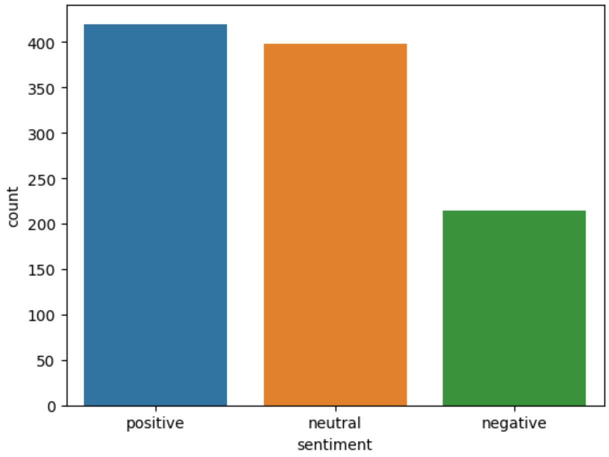


FIGURE 18 A BAR CHART TO SHOW THE SENTIMENT DISTRIBUTION

Figure 19 shows a joint plot for the polarity and subjectivity of the text. The graph is split into 3 parts. The first part is a scatterplot that graphs polarity against subjectivity on a scale of -1 to 1.The colour of the dots indicate whether the sentiment of the word is positive, negative or neutral. The second part is above the x axis and shows the distribution of data for the subjectivity. The third part of the graph is along the y axis and shows the distribution of the data for polarity.

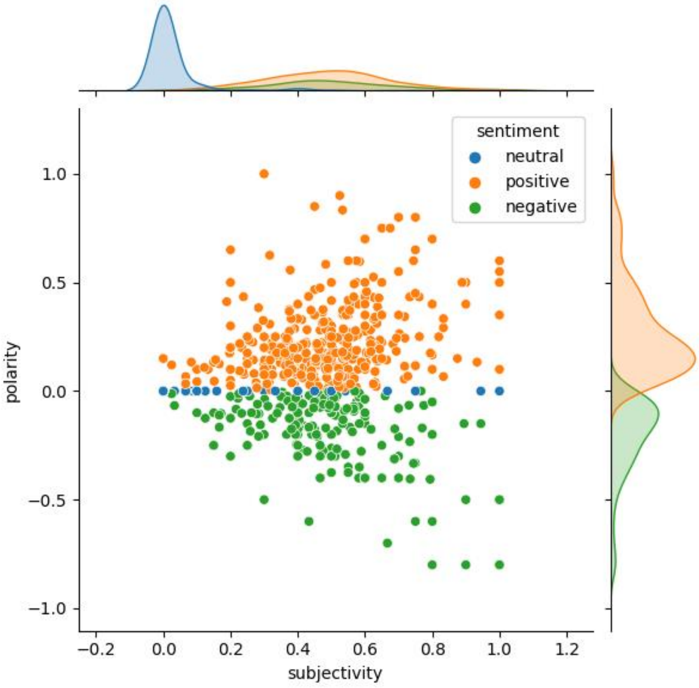


FIGURE 19 A JOINT PLOT TO SHOW THE POLARITY AND SUBJECTIVITY OF EACH WORD

Figure 20 and 21 show the word cloud and frequency distribution for the words with positive sentiment. Figure 21 and 22 show the world cloud and frequency distribution for the words with negative sentiments. Figures 23 and 24 show the word cloud and frequency distribution for the words with neutral sentiments. As show by the visualisations words like 'secur','get','work' and 'job' were among the most used words for all 3 sentiments however the positive sentiments had words like 'good' and 'like whereas the negative sentiments had words like 'threat' and 'shit'.

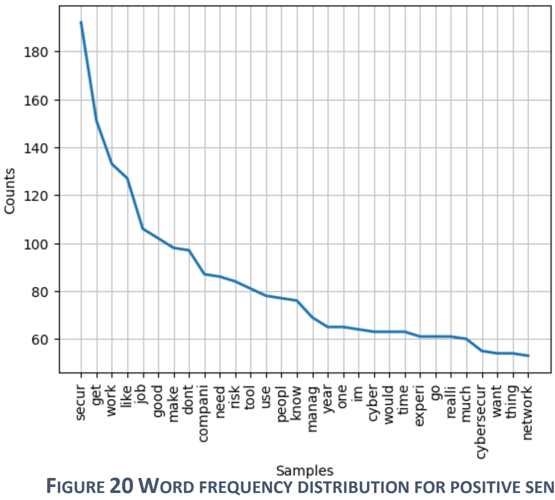


Figure 20 Word frequency distribution for positive sentiment

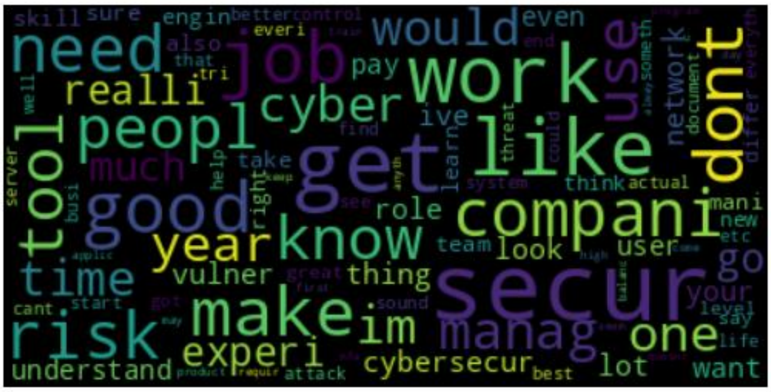


Figure 21 Wordcloud for words with positive SENTIMENT

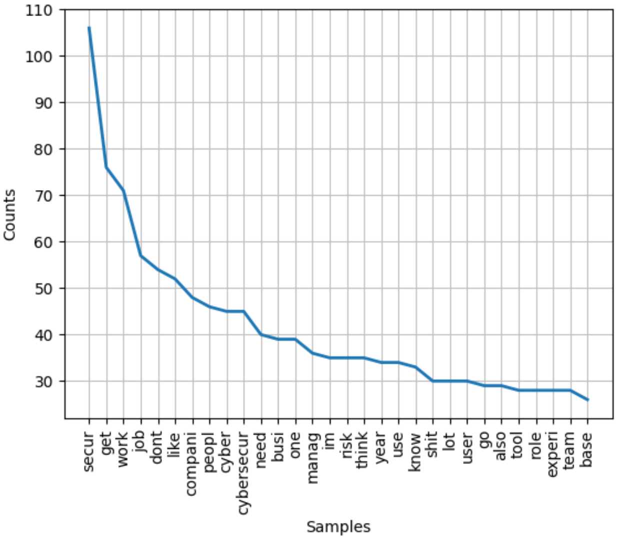


Figure 22 Word frequency distribution for negative sentiment

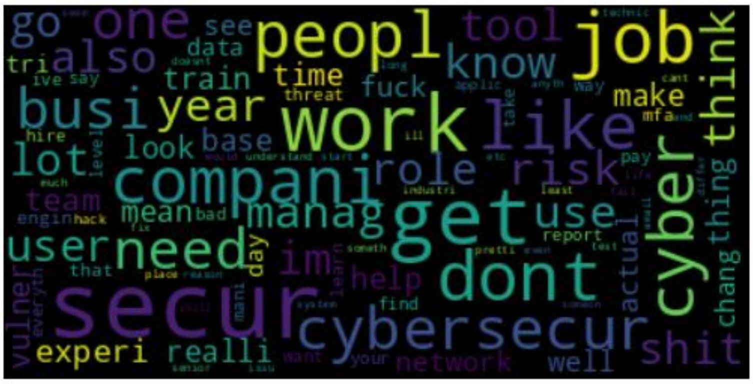
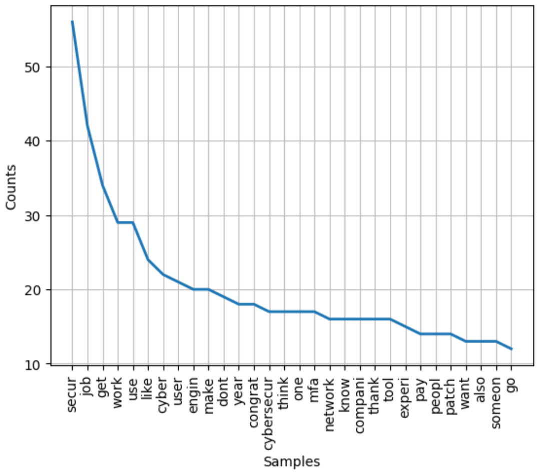


Figure 23 Wordcloud for words with positive sentiment



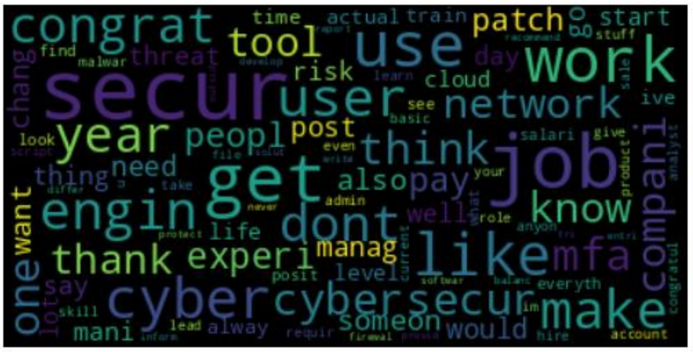


FIGURE 24 WORD FREQUENCY DISTRIBUTION FOR POSITIVE SENTIMENT

FIGURE 25 WORDCLOUD FOR WORDS WITH NEUTRAL SENTIMENT

News API Analysis

News Articles have been pulled from news api for the search term ‘Cost of Living Crisis’. 10 articles were pulled for the purpose of this analysis however Beautiful soup is used as the news Api only allows a maximum of 100 characters to be pulled. Media plays a significant role in society and the circulation of information is very important(Dushyant,2023). The use of news api analysis helps in reducing bias by retrieving news articles from known sources.

Descriptive Analytics

Article	Publish Date	Word Count	Sentences	Characters
Grand National: muted event for UK bookmakers amid cost-of-living crisis	2023-04-17 16:50:53	514	9	3428
Zambians struggle with cost of living as debt rework drags on	2023-04-18 07:10:47	965	25	6318
Tesco not profiteering amid cost of living crisis, says boss	2023-04-13 11:40:11	538	17	3517
Make These Renovations Before Retirement If You Plan to Stay in Your Home	2023-04-17 16:00:00	458	17	2683
UK parents: how has the higher cost of living affected your child maintenance payments?	2023-04-13 00:00:00	192	5	1164
Bank of England expected to raise interest rates again after UK inflation only dips to 10.1% – as it happened	2023-04-19 00:00:00	3647	35	22375
Tell us: how is the UK cost of living affecting your ability to attend weddings?	2023-04-24 00:00:00	247	6	1452
How are English football clubs responding to the cost of living crisis?	2023-04-30 00:00:00	919	35	5680
Shoppers in Great Britain switch to frozen food amid cost of living crisis	2023-04-14 00:00:00	474	9	2883
UK cost of living crisis leading people to gambling, says charity	2023-04-09 00:00:00	644	8	3900

TABLE 7 DESCRIPTIVE ANALYSIS FOR ARTICLES

Average Word count	Average Sentences	Average Characters
859.8	16.6	5340

TABLE 8 AVERAGES FOR ARTICLE DATA

Table 7 shows us the publish date, word count, number of sentences and number of characters for each article pulled from the news API. Table 8 shows the averages for 9 out of the 10 articles as article 6 is being treated as an outlier as the values are considerably higher than the other 9.Figure 26-35 shows the word clouds for each of the articles. Figure 36-45 show the word frequency distribution for each of the articles.

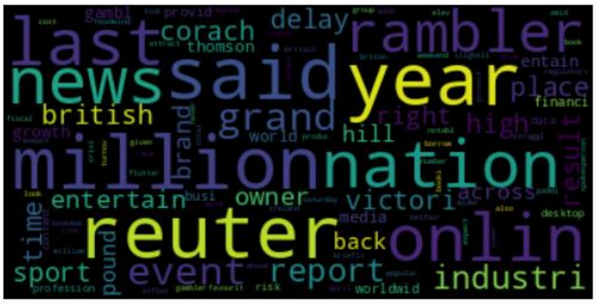


Figure 26 Article 1 wordcloud

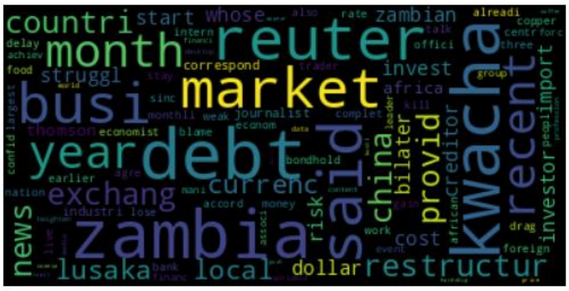


Figure 27 Article 2 wordcloud

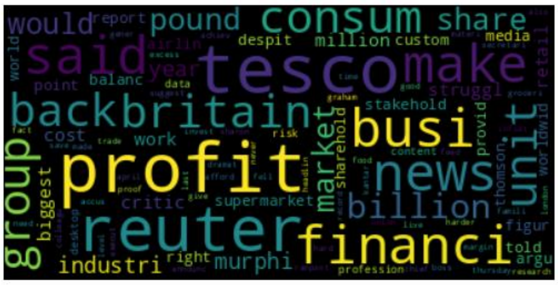


Figure 28 Article 3 wordcloud

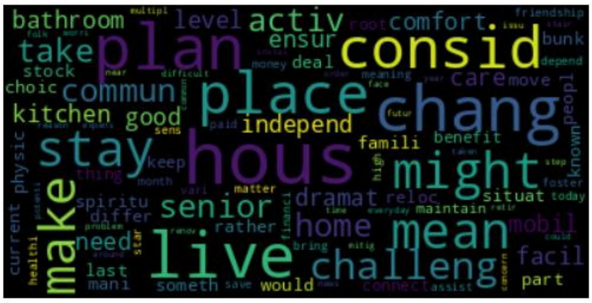


Figure 29 Article 4 wordcloud

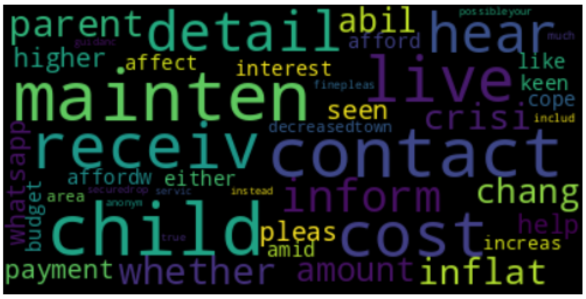


Figure 30 Article 5 wordcloud

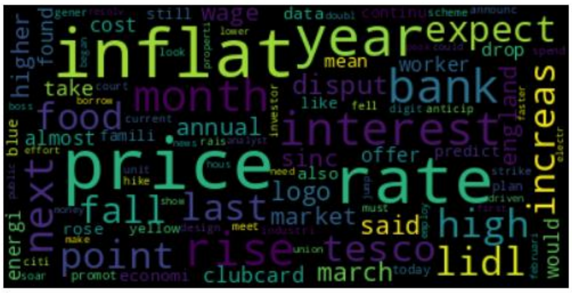


Figure 31 Article 6 wordcloud

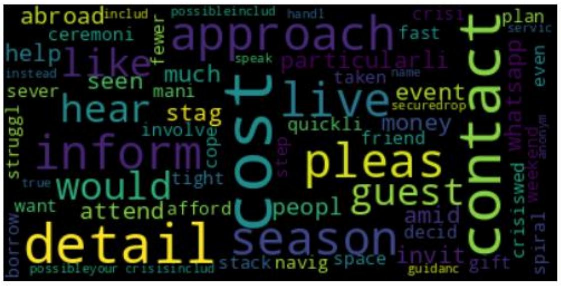


Figure 32 Article 7 wordcloud

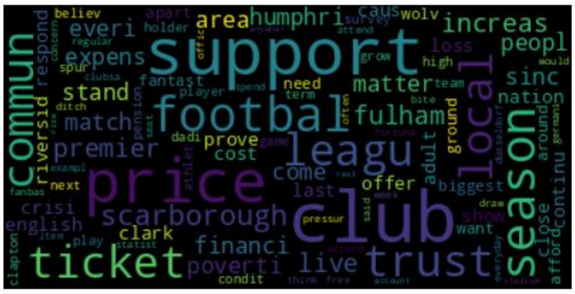


Figure 33 Article 8 wordcloud

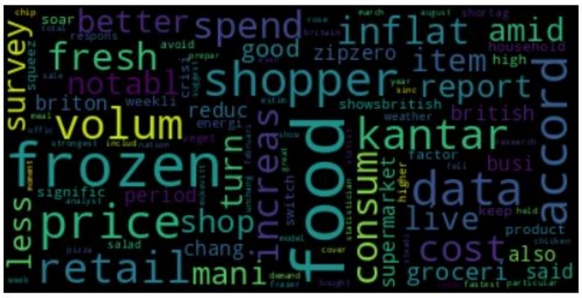


Figure 34 Article 9 wordcloud

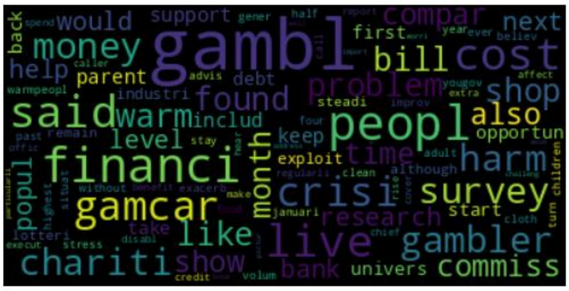


Figure 35 Article 10 wordcloud

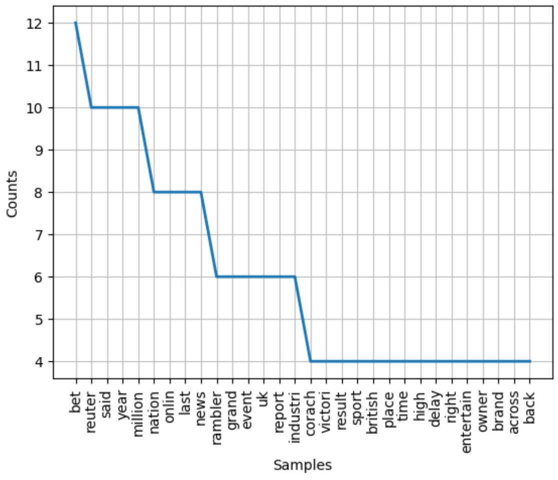


Figure 36Word frequency distribution for article 1

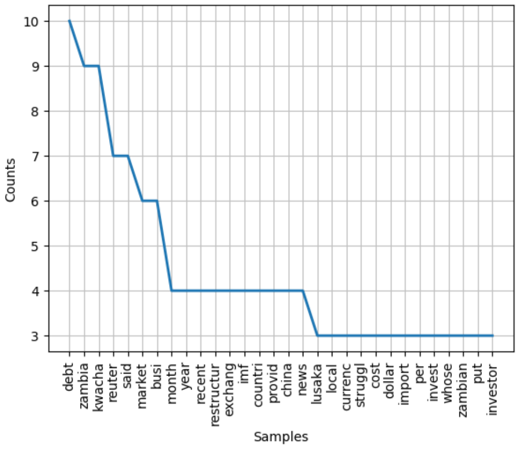


Figure 37Word frequency distribution for article 2

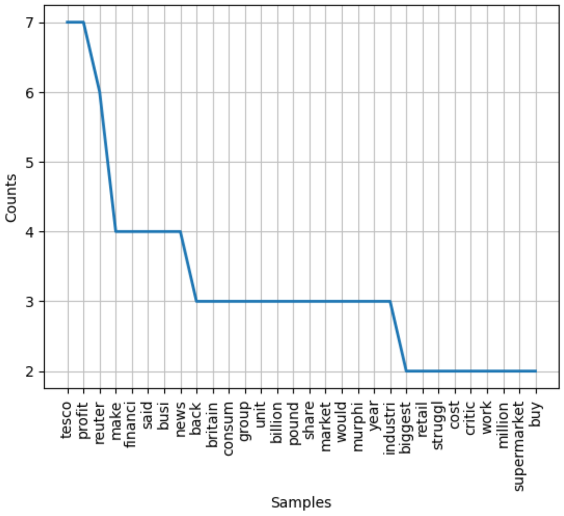


Figure 38 Word frequency distribution for article 3

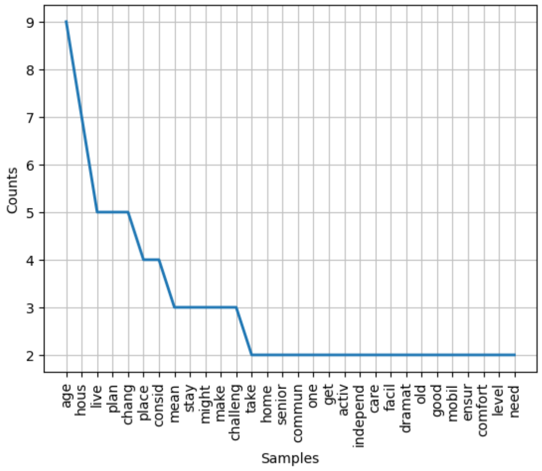


Figure 39Word frequency distribution for article 4

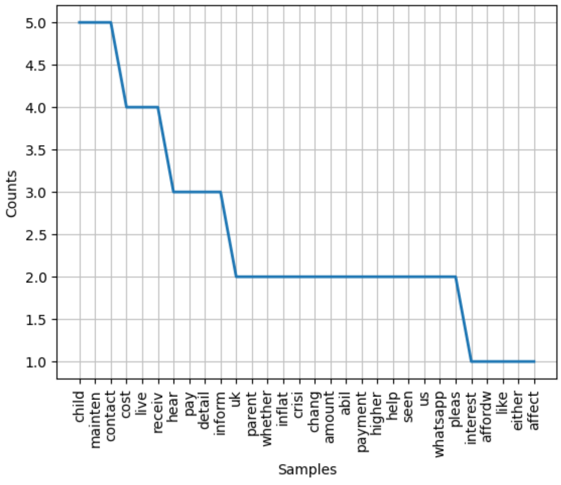


Figure 40 Word frequency distribution for article 5

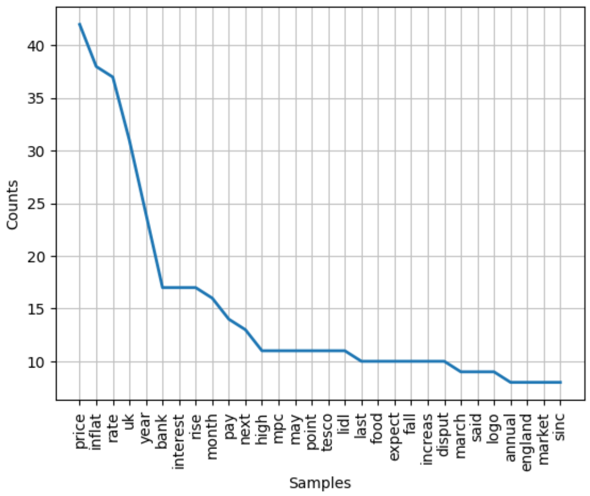


Figure 41Word frequency distribution for article 6

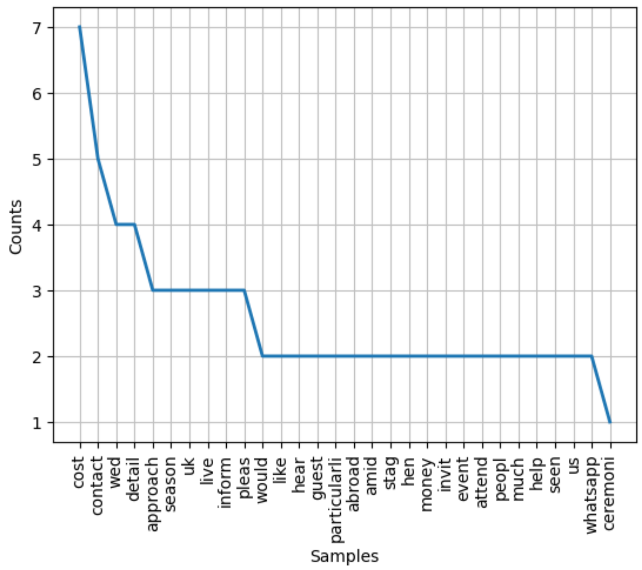


FIGURE 42WORD FREQUENCY DISTRIBUTION FOR ARTICLE 7

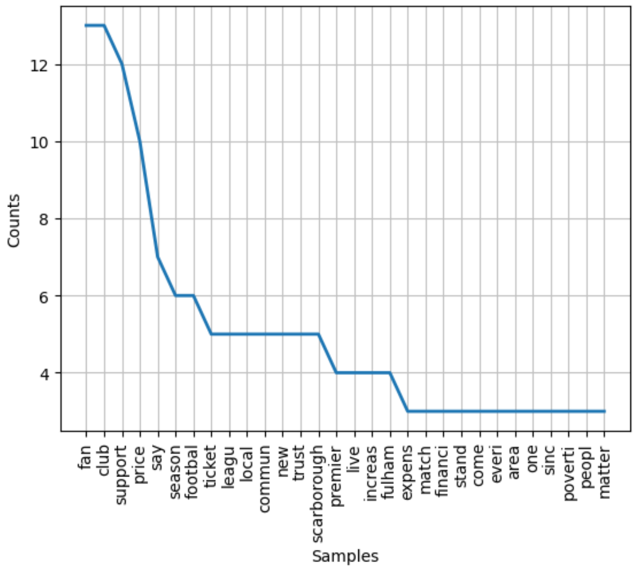
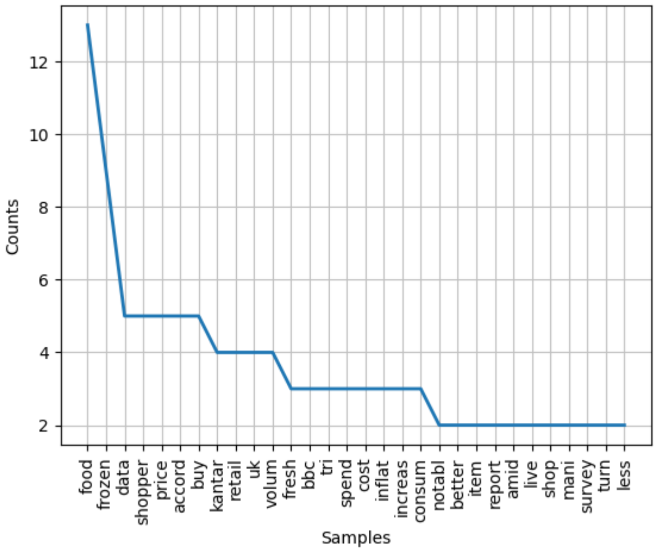


FIGURE 43WORD FREQUENCY DISTRIBUTION FOR ARTICLE 8



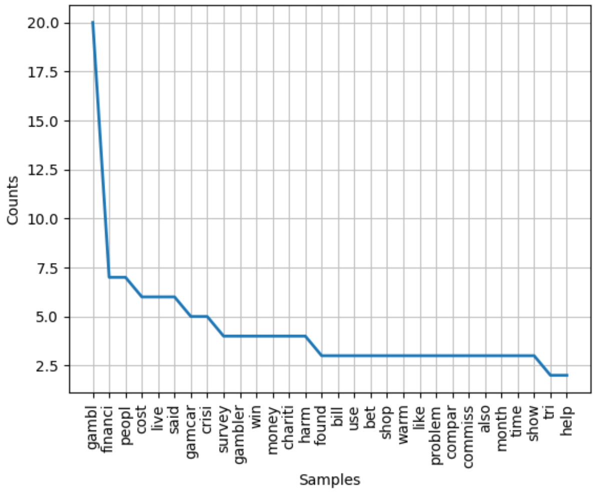


FIGURE 44 WORD FREQUENCY DISTRIBUTION FOR ARTICLE 9

FIGURE 45WORD FREQUENCY DISTRIBUTION FOR ARTICLE 10

Term	Weighting
Inflation	0.355
Prices	0.273
Uk	0.244
Rate	0.191
Year	0.186
Bank	0.164
Interest	0.161
Rates	0.161
Pay	0.136
Next	0.126

TABLE 9 TOPIC 1

Term	Weighting
gambling	0.281
said	0.237
financial	0.200
Reuters	0.199
Cost	0.169
People	0.163
Debt	0.154
Living	0.161
Business	0.136
Kwacha	0.125

TABLE 10 TOPIC 2

Term	Weighting
fans	-0.222
club	-0.173
supporters	-0.173
reuters	0.161
clubs	-0.147
says	-0.144
prices	-0.138
community	-0.128
season	-0.127
scarborough	-0.124

TABLE 11 TOPIC 3

Term	Weighting
gambling	0.481
fans	-0.143
living	0.127
cost	0.126
reuters	-0.118
people	0.116
supporters	-0.111
club	0.111
crisis	-0.103
gamcare	-0.102

TABLE 12 TOPIC 4

Term	Weighting
food	-0.437
frozen	-0.355
shoppers	-0.196
buying	-0.191
according	-0.188
data	-0.170
kantar	-0.160
retail	-0.158
gambling	0.141
fresh	-0.119

TABLE 13 TOPIC 5

Term	Weighting
debt	0.214
kwacha	0.185
tesco	-0.171
reuters	-0.141
zambias	0.139
million	-0.136
news	-0.127
last	-0.114
britains	-0.112
online	-0.110

TABLE 14 TOPIC 6

Term 5	Weighting
house	-0.330
changes	-0.237
age	-0.233
you’re	-0.201
you’ve	-0.197
Consider	-0.190
place	-0.169
means	-0.143
aging	-0.142
might	-0.142

TABLE 15 TOPIC 7

Term 5	Weighting
contact	0.306
cost	0.248
information	0.229
living	0.205
maintenance	0.196
child	0.195
hear	0.192
please	-0.190
receive	0.157
helpful	-0.153

TABLE 16 TOPIC 8

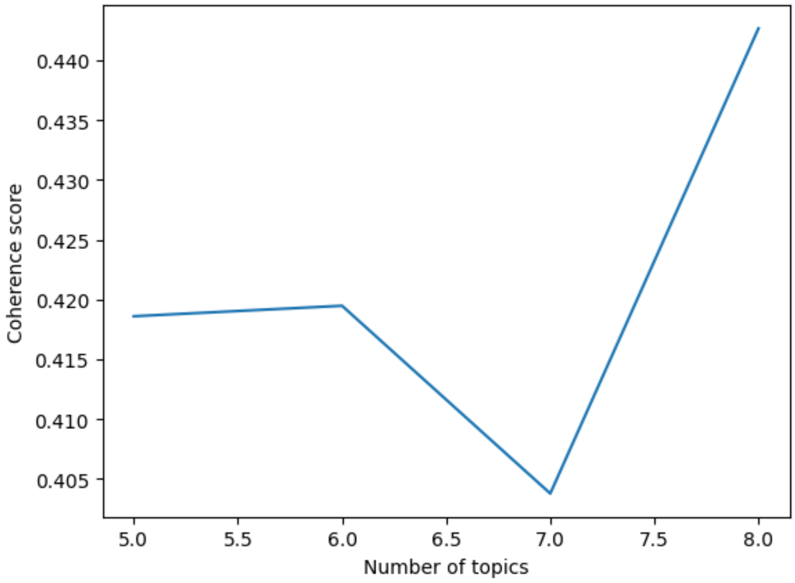


FIGURE 46 A GRAPH TO SHOW COHERENCE BY NUMBER OF TOPICS

Topic Modelling

Figure 46 shows the coherence score by number of topics.As shown by the graph the coherence starts going down at 6 topics where it hits the lowest at 7 topics.It then starts going up at hits it’s peak at 8 topics. The range for topics is 5 -8 topics but as shown by the graph it is projected to go up further so in the future going higher could be beneficial. Table 9-16 shows the weighting for each term for the 8 topics. As is visible there are a few overlapping terms like ‘cost’,’prices’ and ‘people’.

Article Summarisation

The article chosen for summarisation is an article by The Guardian about the switch from fresh food to frozen food amidst the cost-of-living crisis. To begin the word frequency distribution is calculated. After this the sentence importance is calculated using sentence scores. The average sentence score is then computed and if the sentence score is 1.5 times bigger than the average sentence score then the sentence is added to the summary.

The summary has areas in which the summary which is good like where it correctly quotes Mohsin Rashid however one of the limitations is that it also begins with the quote but the quote at the beginning doesn't make sense. It does go into some of the important parts of the article. Overall it's a good summary but by no means perfect.

Summarised text of article

"And some of that is clearly to do with the cost of living," he added.A quarter of UK shoppers say they are buying more frozen food, according to a separate survey of 2,000 British adults by the pollster Opinium on behalf of Zipzero, an app that collects shoppers' receipts data in exchange for cash.The poll also found that 30% of people are buying more food from the reduced section of supermarkets to try to save money, and 21% are buying less meat and fish.Sign up to Business Today.Get set for the working day – we'll point you to all the business news and analysis you need every morning after newsletter promotion.Mohsin Rashid, Zipzero's chief executive, said: "Sky-high food inflation has invariably shifted consumer habits.

Conclusion

Overall this report has focused on some key analysis on the posts from the biggest tech subreddits. It has also looked at sentiment analysis for the comments on the cybersecurity subreddit. Following that it has looked at a twitch network graph and looked at degree,betweenness and eigenvector centrality.To finish the News api was used to pull 10 articles of which visualisations were provided and topic modelling was conducted. Going forward more topics could be explored as well as using other models to get coherence score. Machine learning could also be applied to predict many other things and provide more insights.

References

Dive into anything (no date) *Reddit*. Available at: https://www.reddit.com/r/computing/comments/2e38m6/master_list_of_tech_subreddits/ (Accessed: May 5, 2023).

Dushyant (2023) *News API significance and use cases in different industries -*, *Newsdata.io*. Available at: https://newsdata.io/blog/news-api-significance/ (Accessed: May 5, 2023).

Golbeck, J. (2013) "Network structure and measures," *Analyzing the Social Web*, pp. 25–44. Available at: https://doi.org/10.1016/b978-0-12-405531-5.00003-1.

Perez, C. and Germon, R. (2016) "Graph creation and analysis for linking actors: Application to Social Data," *Automating Open Source Intelligence*, pp. 103–129. Available at: https://doi.org/10.1016/b978-0-12-802916-9.00007-5.

shaw, alan (2019) *Understanding the concepts of eigenvector centrality and Pagerank*, *Strategic Planet*. Available at: https://www.strategic-planet.com/2019/07/understanding-the-concepts-of-eigenvector-centrality-and-pagerank/ (Accessed: May 5, 2023).

Stafford, C. (2016) *What is reddit?: Definition from TechTarget*, *CIO*. TechTarget. Available at: https://www.techtarget.com/searchcio/definition/Reddit (Accessed: May 5, 2023).