Final Report: Web Social Media and Analytic

Muhammad Ammar Shahid (23103156) * [1]

Word Count: 1820

1. Introduction

Today two third majority of internet users use social media. According to (Ortiz-Ospina & Roser, 2023) there were 7.7 billion people on this planet earth, out of which 3.5 billion use the internet, and this number is just increasing. The authors discussed the rise in the use of social media. Now a days business executives use social media for data driven decisions to boost profits. However, the term social media itself is broad, to be clear and know the type of platform that is most suitable for individual, or an organization (Kaplan & Haenlein, 2010) discussed some critical points. Twitter is a platform where users can share their thoughts and opinions. Sentiment analysis in twitter deals with the problem of analyzing tweets for which express the opinion of users.

The data itself is passive and tells nothing unless someone gets the insights from it. This report will analyze the twitter data for the political situation in Pakistan after the dismissal of the Imran khan's government to know the user opinion. Furthermore, will perform some topic modeling techniques on the news articles and summarize the article. At last report will elaborate some techniques to analyze the graphs and will get some insights from a graph data-set. The focus of this report is to implement all the techniques that are studied in the module.

2. Twitter Data Analysis

2.1. Data Collection

Twitter data was collected to interpret the user opinions on the dismissal of Imran khan's government. Total 20,000 tweets which were containing the words Regime change and imported government (10,000 each) were collected, out of which around 18,833 were in English formatted text, so rest of tweets were screened out. Fig 1 elaborates the data columns of the scraped data and Fig 2 gives an overview of what data looks like. Text column contains the tweets.

```
Int64Index: 18834 entries, 0 to 18833
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Datetime  18834 non-null  object
 1   Tweet Id  18834 non-null  float64
 2   Text      18834 non-null  object
 3   Username  18834 non-null  object
 4   Hashtag   5648 non-null   object
 5   Views     4498 non-null   float64
 6   Retweet   18834 non-null  int64
 7   Place     410 non-null    object
 8   Lang      18834 non-null  object
 9   Source    18834 non-null  object
```

Figure 1. Data Headers

	Datetime	Tweet Id	Text	Username	Hashtag	Views	Retweet	Place	Lang	Source
0	2023-02-28 22:22:22+00:00	1.630690e+18	@WilliamRHawkins @molotov2_5 @DanielLMcAdams @...	chinoslims	NaN	34.0	0	NaN	en	Twitter for Android
1	2023-02-28 20:41:57+00:00	1.630670e+18	Downward spiral of regime change continues to ...	kamranalishah	NaN	24.0	0	NaN	en	Twitter for Android
2	2023-02-28 19:24:42+00:00	1.630650e+18	Imran Khan Government was the most public frie...	FactCheckAsia	NaN	603.0	12	NaN	en	Web App
3	2023-02-28 13:15:11+00:00	1.630560e+18	@HniaziISF Hassan what is the method to remove...	sandyskye22	NaN	70.0	0	NaN	en	Twitter for iPhone
4	2023-02-28 11:10:54+00:00	1.630530e+18	The reprehensible arrest of Amjad Shoaib shows...	Shahidparvezak2	['Ø®Ø§Ù_Ø¨ÙÙØ§Ø¨ÙÛ_Ú©Ø±Ù¾Ù¹_ÛØ¸Ø§Ù']	611.0	66	NaN	en	Twitter for iPhone

Figure 2. Data Rows

2.2. Data Pre-processing

Collected data was lacking the inconsistency, for instance date and time needs to be formatted, similarly; data in sources was in the form of long strings, also in place column city and country both were mentioned to analyze it better cities were removed.

2.3. Exploratory Data Analysis

2.3.1. Users with maximum tweets

User with unique usernames were counted in the data to know the number of times they tweeted. Hence, counted usernames were sorted in descending order to know about the user with maximum number of tweets. Fig 3 illustrates that user with username adamjeezee tweeted most with 276 tweets, and the list continues in the descending order.

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: <muhammad.shahid3@mail.bcu.ac.uk>.

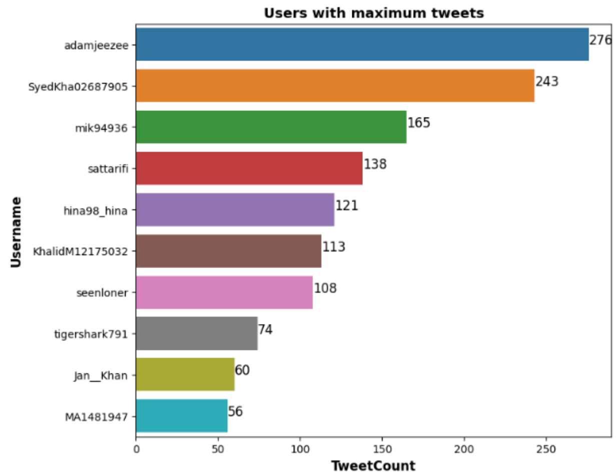

Figure 3. Most tweets bar plot

2.3.2. Users with highest views

Fig 4 shows the users with highest views. With this kind of analysis it can be presumed that user with higher views has more following. AVeteran1956 is more popular user amongst others, and it has more than 140,000 views on his tweets.

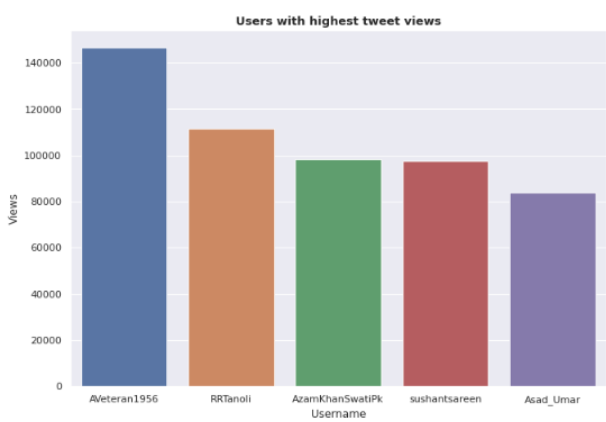

Figure 4. Highest views bar plot

2.3.3. Retweet with respect to location

Out of 18,831 tweets only 410 had location tag, which clearly shows that users are very much concerned about their locality. Based on available data Fig 5 elaborates that users from Pakistan has higher number of tweets, which shows that they are more expressive about the topic which is described in 1.

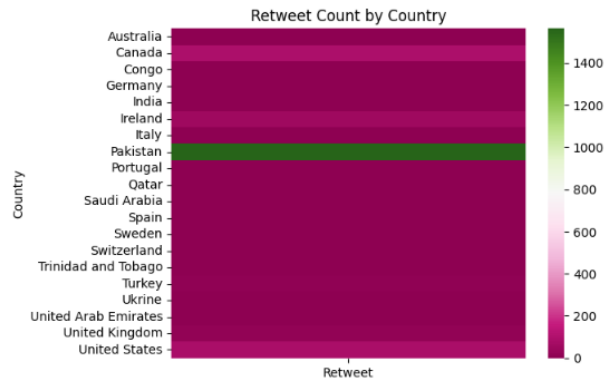

Figure 5. Retweets with respect to place (Heatmap)

2.3.4. Source of tweets

Twitter is a platform that can be accessed from any type of device, to know the most popular source of tweets amongst the users in data; analysis was run on the source column and Android appears to be the most used device for tweets 6 defends this argument by showing that 59 percent of users use android phones.

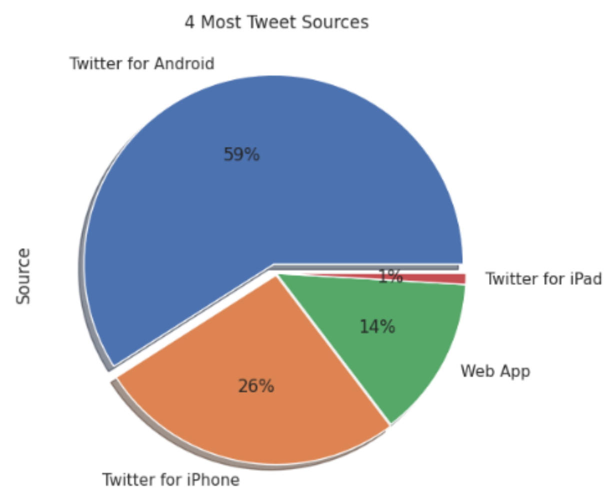

Figure 6. Pie chart for Source of tweets

2.4. Sentiment Analysis

2.4.1. Tweet cleaning

Raw tweets were containing so much noise in the form of unwanted text user-tags and emojis as shown in 7. So before performing any further analysis, unwanted words, punctuation and characters were removed.

Figure 7. World Cloud for Raw text

2.4.2. POLARITY AND SUBJECTIVITY

After text cleaning, next step was to convert the tweets into words to apply lemmitization or stemming to change the words into verbs or root form respectively. TextBlob (library in python) was then used to check the polarity and subjectivity of each tweet. Tweets with 0 polarity were labelled as neutral, similarly; tweets with the polarity less than 0 and greater than 0 were categorized as negative and positive respectively. Fig 9 justifies that positive tweets are more in the data. However, there is fractional difference between positive and negative tweets. Fig 8 discuss about both the polarity and subjectivity and elaborates that neutral tweets have more subjectivity as compared to positive and negative tweets.

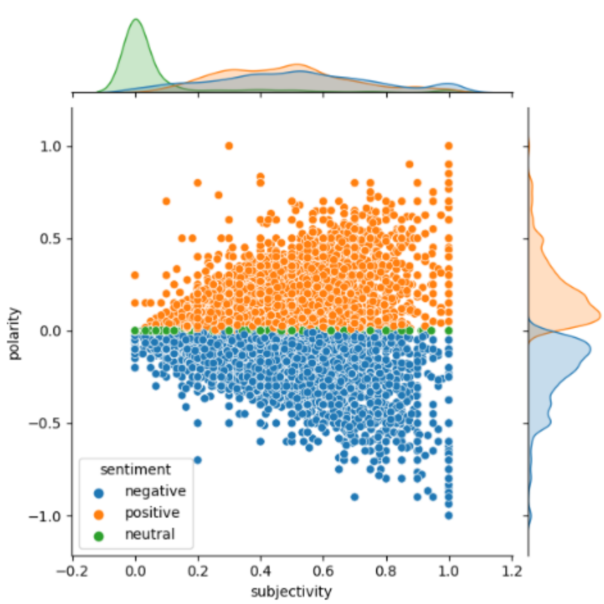

Figure 8. Polarity and subjectivity of tweets

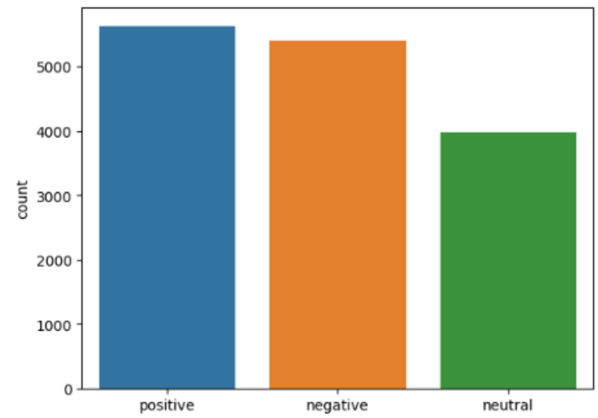

Figure 9. Sentiment distribution of tweets

2.4.3. WORD FREQUENCY AND WORD CLOUDS

This section will observe the positive, negative and neutral words in the tweets. Although Word clouds are self explanatory, in positive tweets users are talking about right leader, operation, and elections etc see Fig 10 and Fig 11, whereas in negative tweets users are discussing corruption, criminals and PDM etc see Fig 12 and Fig 13. Some words are repeating in all sentiments (positive and negative) this is because of the contextual meaning.

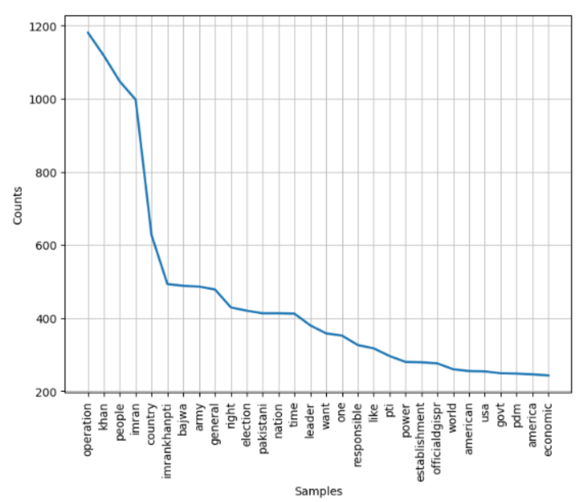

Figure 10. Positive word frequency

2.5. Automation for Future Predictions

Data was collected, cleaned, and analyzed. The next step was to get most out of useful insight; so, instead of destroying the data tweets and their polarity were saved in a file as independent (tweet text) and dependant (polarity) as shown

Figure 11. positive word cloud

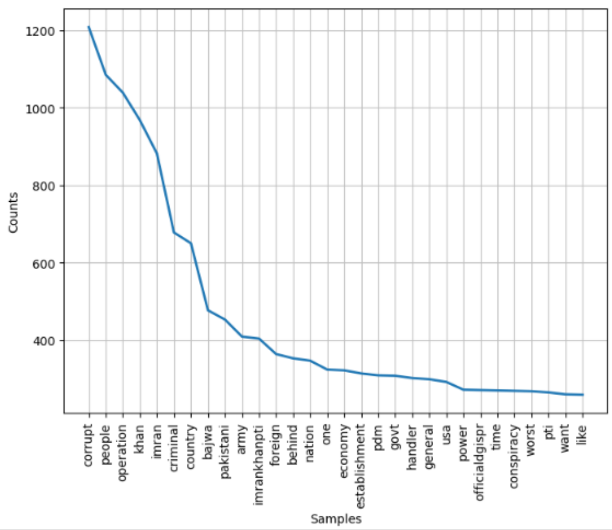

Figure 12. negative word frequency

in Fig 14; only positive and negative tweets were captured, with the help of this data a neural network was trained to predict the any future tweets. This made easy to analyze the sentiment of a tweet.

Machine learning (ML) was trained using TensorFlow frame work on around 13,000 tweet data. Data was split into 70 percent for training and 30 percent for testing. Model parameters are described in Tab 1 and it was trained for 12 epochs using the batch size of 80. On the testing data model achieved the accuracy of 88 percent.To see the accuracy graph and confusion matrix refer to Fig 15 and Fig 16 respectively.

3. News Article Analysis

Data was collected using News-Api to analyze the articles. Basically Four topics were searched, including **Regime Change**, **Regime Change in Pakistan**, **Imported Government** and **Elections in Pakistan**. Articles on all four topics were saved into a single corpus. After removing HTML tags and data cleaning word frequency was counted to see which words are most frequently appears in the corpus see Fig 18, and then word cloud was build to get the insight of the text data. Word cloud showed the most relevant information from the text see Fig 17

Figure 13. negative word cloud

Table 1. parameters for neural network

Layers	3
Dropout	0.5
Activation function	Relu and Sigmoid
Optimizer	RMSprop
Loss function	Binary cross-entropy

3.1. Topic Modeling

After performing word analysis, next step was to choose the number of topics for the four articles' combined corpus. For that purpose different number of topics were tried starting from 1 to 12. Coherence was also measured to check the best number of topics that can be given to the corpus. Figure 19 elaborates the coherence against every number of topic in a line graph, and graph shows that 3 number of topics are best for the corpus. As there were four topics in total and two of them were almost same (Regime Change and

	Text	Text_polarity
0	williamrhawkins daniellmcadams slavyangrad off...	0.0
1	downward spiral regime change continues hurt c...	1.0
2	imran khan government public friendly governme...	1.0
3	hniaziisf hassan method remove corrupt judge p...	0.0
4	reprehensible arrest amjad shoaib show democra...	1.0

Figure 14. Data head for training

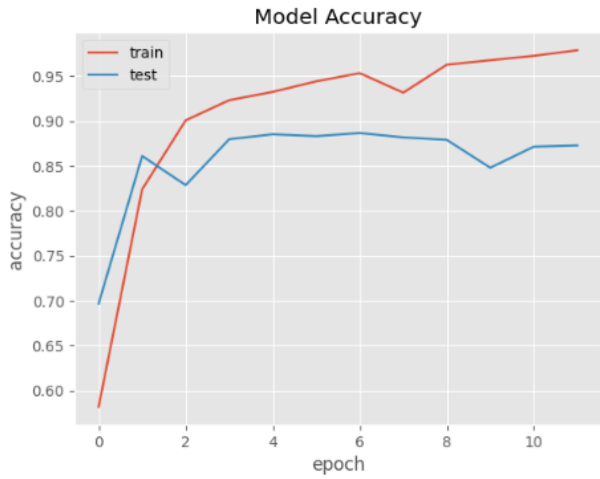

Figure 15. Accuracy achieved during the epochs

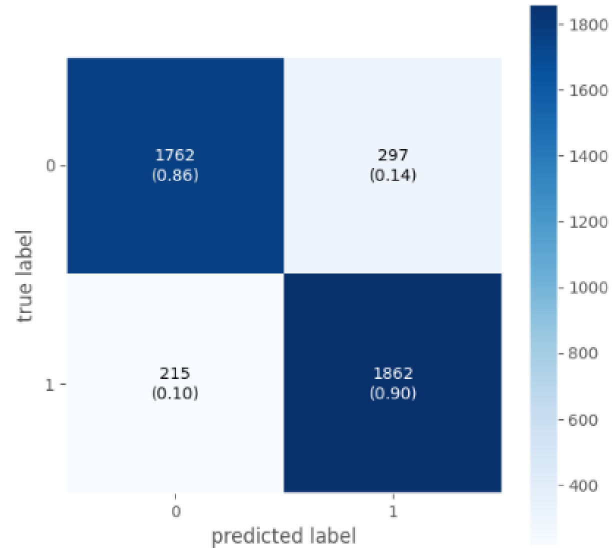

Figure 16. Confusion matrix

Regime change in Pakistan) so that could be the why three numbers of topics got highest coherence.

3.2. Article Summary

This section will discuss about the summary produced after analysis the sentence importance. One article (Regime change) was chosen for this task. Based on the word frequency distribution importance of every sentence was calculated and then the average of all sentences was calculated. Sentences with the importance more than 3 times of the average were choose to make the summary of the corpus. Summary of the Article is as follows;

The key stakeholder and people's expectationsPeaking inflationOne wonders why regime change took place in Pakistan and how internal crisis deepened after Imran Khan's removalHistorically, no Pakistani Prime Minister has completed a five-year term — assassinated, hanged, or removed through the imposition of martial law or by the president under the 8th constitutional amendment. Despite former Prime Minister Imran Khan's allegations of a foreign conspiracy, this time the method for regime change was constitutional.Britannica defines regime change as "overthrowing a government considered illegitimate by an external force and its replacement with a new government according to the ideas or interests promoted by that force." Cambridge dictionary defines regime change as "a complete change of government, especially one brought about by force."In Western democracies, elected governments are removed through a vote of no-confidence or popular protests, which compel the party in power to replace the head of government or state through constitutional means. On the other hand, Moscow has tried to have the pro-West Ukrainian President Volodymyr Zelensky removed.One cannot help but wonder why regime change took place in Pakistan and how internal crisis deepened after Imran Khan's removal.

Figure 17. News articles word cloud

4. Graph Analysis

The graph data (EU email communication network) was taken from (Leskovec et al., 2007). It's a network of email data of an European research institute. It's a record of email sent and received between the staff. In total there were 265214 nodes and 420045 edges, diameter of the graph was 14 but as the data was quite big and requires more computational power and time; so, only first 80,000 records were taken for the analysis. In the subset of the data there were 41,296 nodes and 72,799 edges.

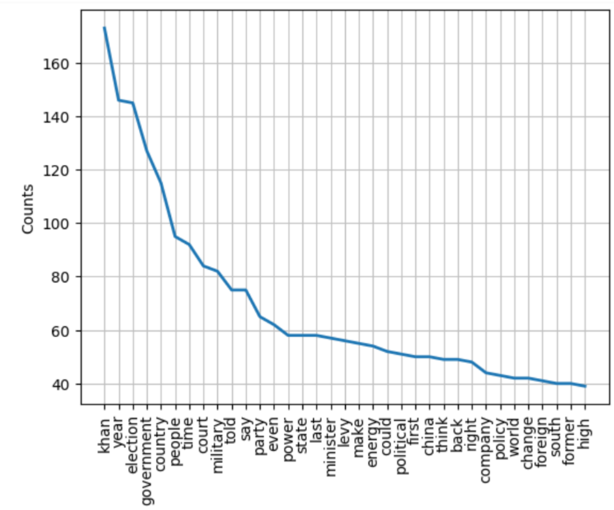

Figure 18. News articles word frequency

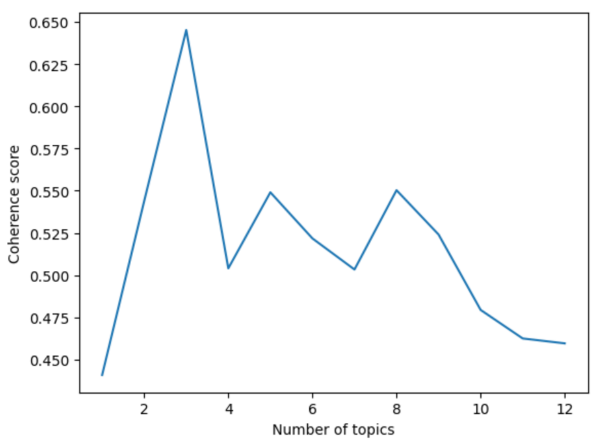

Figure 19. Number of topics and their coherence

4.1. Degree centrality and normalized Degree Centrality

Not every node in the graph is important, but importance can be measured based on several measures. One of them is Degree, and it can be defined as number of nodes attach to a certain node is it's degree, that means more the neighboring nodes the more important node is. On the other hand, normalized degree centrality considers the relative importance of a node's degree centrality compared to other nodes in the graph. It re-scales the degree centrality values to a specific range, often between 0 and 1, to create a standardized measure of centrality that enables meaningful comparisons across graphs of varying sizes and structures. Fig 20 is for Degree centrality and Fig 21 is for normalized degree centrality. Average degree of a node in the graph is 3.52, that explains that on average each node is connected to 3 other nodes in the graph.

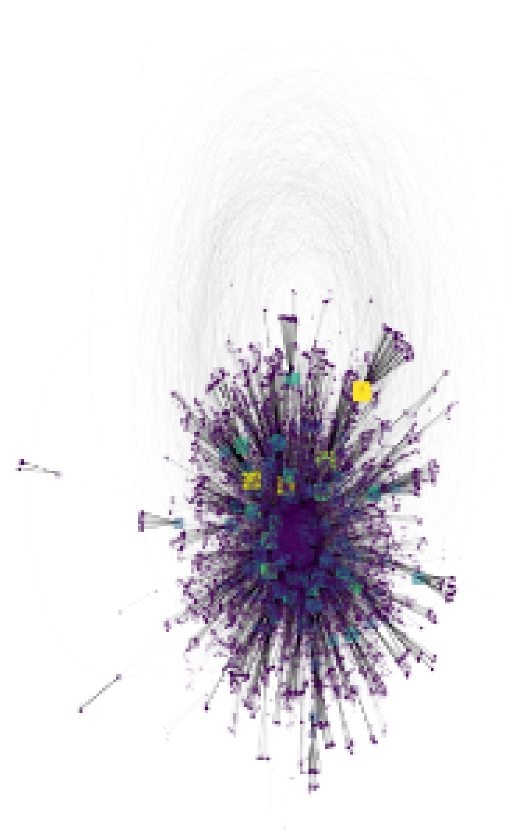

Figure 20. Degree centrality

4.2. Betweenness Centrality

In this type of centrality importance of a node is measured by the number of the shortest paths passes through a specific node, higher the number higher the centrality would be. Fig 22 refers to the graph for the betweenness centrality. In the graph bigger nodes represents high centrality.

4.3. Eigenvector Centrality

It calculates the centrality based on the influence of the node. To explain further connection of a node to other influential nodes is calculated and if node is connected to important nodes its centrality would be high. Fig 23 illustrates the influential nodes in the data

4.4. Clustering Coefficient

It represents that how strong a community is, in other words how well connected the community is. High value of clustering coefficient (CC) represents higher connectivity with in the nodes. Usually value of average CC value lies be-

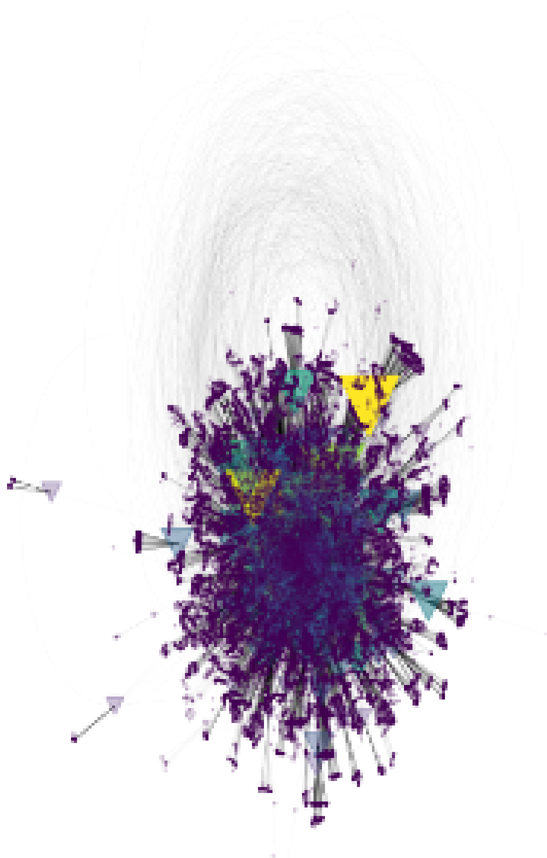

Figure 21. Normalized degree centrality

tween 0 and 1 and value ¿ 0.50 is considered to be the strong. Average CC value for this data set is 0.11 which means EU researchers' community was not strong.

References

Kaplan, A. M. and Haenlein, M. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2009.09.003. URL https://www.sciencedirect.com/science/article/pii/S0007681309001232.

Leskovec, J., Kleinberg, J., and Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1 (1):2–es, 2007.

Ortiz-Ospina, E. and Roser, M. The rise of social media. *Our world in data*, 2023.

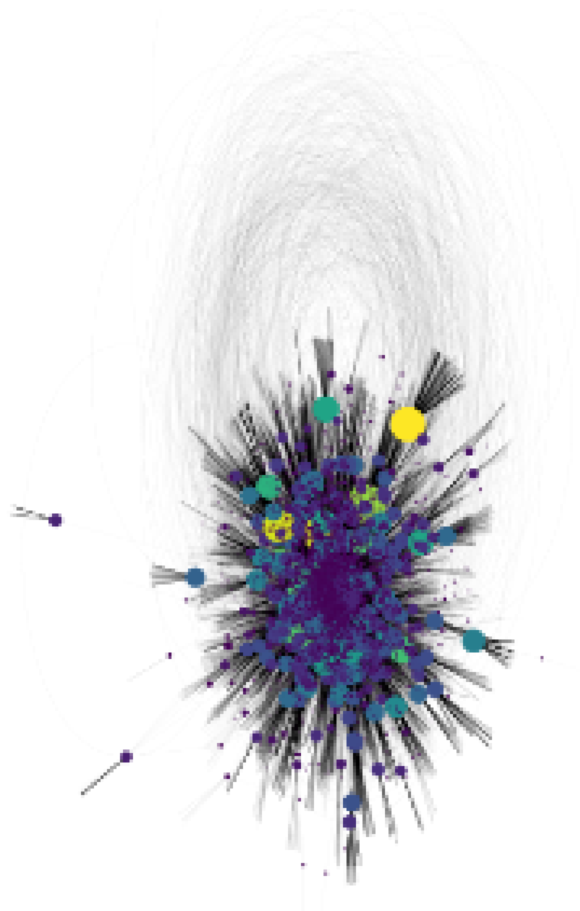

Figure 22. Betweenness centrality

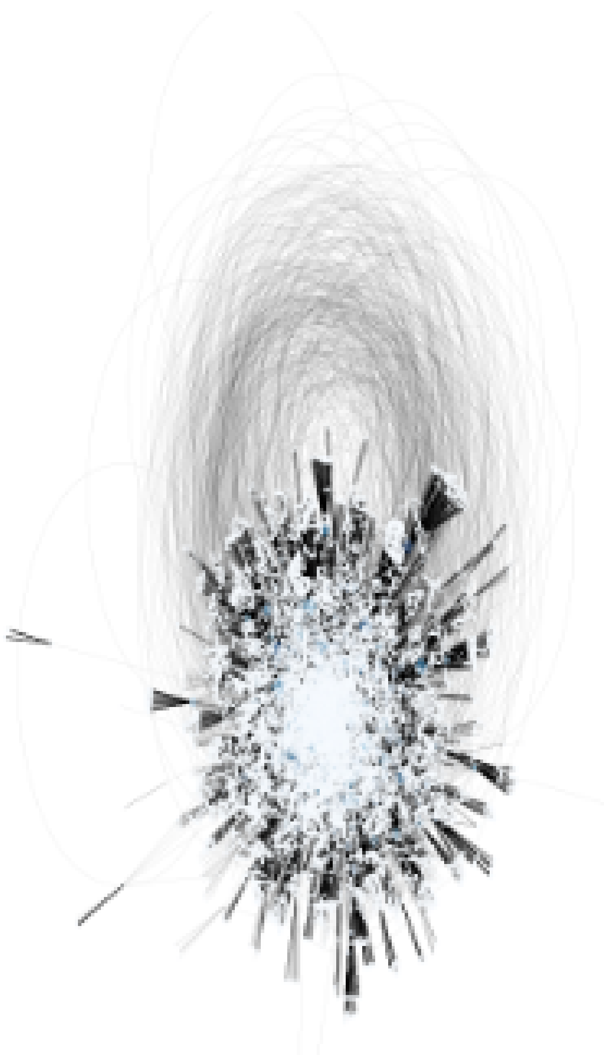

Figure 23. eigenvector centrality