Social Network Analysis of 2020 USA Presidential Election Tweets

Jayanth Anala
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
jayanth.srivaastava2019@vitstudent.ac.in

Dhanush Reddy Pothala
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
pothaladhanush.kumar2019@vitstudent.ac.in

Manish Chembeti
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
chembeti.manish2019@vitstudent.ac.in

2022 International Conference on Advances in Computing, Communication and Materials (ICACCM) | 978-1-6654-7439-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACCM56405.2022.10009172

***Abstract*—Internet surfers and smartphone owners frequently utilize Twitter applications to post their opinions and ideas on different things happening around the world. In recent studies, it can be seen that these posts or tweets can have a major impact on real-life situations and scenarios. The impact twitter has on society is what motivated us to conduct this analysis. This paper is concerned with carrying out the best analysis of tweets based on the US presidential elections that happened in the year 2020. Approximately 5000 tweets from key candidates for the 2020 US presidential election were subjected to network analysis. We determined the election campaign phases and the number of times candidates mentioned this issue on Twitter. The leader-follower connections between the candidates were visualised in this article. We get a conclusion on which Twitter account for campaigns performs better. We'll use the different centrality measures in this paper such as Betweenness, Closeness, Eigen Vector, degree centralities and modularity. As previously said, we'll use a few metrics to compare different results obtained.**

***Index Terms*—Social Network Analysis, Centrality, Community Detection, Zephi Tool, Graphs**

I. Introduction

Looking at recent tweets that were retweeted by each contender and utilising their campaign hashtags, we want to determine the US Presidential candidate and their tweets that are exhibiting more influence and connecting with general people. Items with the "#" sign are known as hashtags. When a user of a social media network posts something relating to a trend that is now trending on the network, they are used. The tweet that is more interesting may be determined by using these hashtags and evaluating them. This is so that Twitter Inc. may keep all postings that use the same hashtag in the same database. The power of social media as a source of interesting news outlets has been demonstrated. Recent analyses show that in 2016, Twitter had a significant influence on the campaign. It evolved into a formal forum for candidates to submit the ideas and actions they intend to do, which instantly and directly reach users' hands.

Nowadays, Twitter marketing is a useful tool for campaigns. The most important platform for political advocacy and boosting voter participation is Twitter. This platform is a really powerful tool for spreading your message. Additionally, because Twitter moves so quickly, it is simple for politicians to slip up or lose contact with their audience. There are several ways and strategies for successful Twitter campaigning like Hashtags, PR strategies, interesting posts and many more. Out of which Hashtag has more rate of success.

The main objective of our project is to find which of either president's campaigns is showing more impact on social networks and handles. We are considering Twitter over Facebook as they made it easy to access their data using public API. We use the betweenness centrality measure to obtain the required result from the network we form using the above data. The Network we form is based on whether a tweet is being retweeted or not. We can forecast the outcome of the forthcoming elections using this finding.

II. Related Work

Measures of centrality are an essential tool for comprehending and network analysis, also known as graph analysis. To determine the significance of any specific node in a network, these techniques make use of graph theory. They show areas of the network that require attention by cutting through noisy data. We need to comprehend how different centrality metrics operate for visualisation applications because they each have their own meaning of "important".

*A. **Betweenness Centrality***

Betweenness centrality, which is based on shortest routes, is a measure of centrality in a graph in graph theory. In a connected graph, there always presents at least one shortest route between any pair of vertices such that the total of the edge weights or the number of edges the path travels or passes through is reduced. In network theory, betweenness centrality

The source code and datasets used for this analysis can be found on github, contact author.

plays a vital role, it shows how far apart the nodes are from one another. A node with a greater or higher betweenness centrality in a communications network would have greater control over the network since more information will travel through that particular node. It is applicable to a broad variety of network theory issues, including issues with social networking, biology, transportation, and collaborative research.

$$\mathrm{g}(\mathcal{V}) = \sum_{\mathrm{p}\neq\mathrm{v}\neq\mathrm{q}} \frac{\sigma_{\mathrm{pq}}(\mathrm{v})}{\sigma_{\mathrm{pq}}} \tag{1}$$

σ_{pq}, is the quantity shortest paths between q and p.
$\sigma_{pq}(v)$, is the quantity of those routes that are passing via v.

B. Degree Centrality

In network theory, degree centrality is said to be one of the simplest topics, which is said as the number of linkages occurring at a node (the total number of links that a node has). One particular node's immediate danger of catching whatever is going across the network can be expressed in terms of the degree. We often define two distinct measures of degree centrality, referred to as in-degree and out-degree, in the context of a directed network. A node's in-degree counts the links that are directed toward it, while its out-degree counts the ties it sends out to other nodes.

If the degree centrality of the vertex is A', of a graph G=(A, S) with S edges and A vertices, is defined as,

$$\mathsf{C}_{\mathrm{D}}(\mathcal{A}) = \deg(\mathcal{A}) \tag{2}$$

For suppose, A* is the node which has the more degree of centrality in a respective graph P, let X=(U, Z), which increases the following quantity for a U node connected to a graph. U* with the node which has the more degree of centrality in X.

$$\mathsf{H} = \sum_{\mathrm{j}=1}^{|\mathrm{U}|}[\mathbf{C}_{\mathrm{D}}(\mathrm{U}*) - \mathbf{C}_{\mathrm{D}}(\mathbf{U}_{\mathrm{j}})] \tag{3}$$

Degree centralization, for same graph P,

$$\mathsf{C}_{\mathrm{D}}(\mathrm{P}) = \frac{\sum_{\mathrm{i}=1}^{|\mathrm{A}|}[\mathsf{C}_{\mathrm{D}}(\mathrm{A}*) - \mathsf{C}_{\mathrm{D}}(\mathrm{A}_{\mathrm{i}})]}{\mathrm{H}} \tag{4}$$

C. Closeness Centrality

The following general equation in two variables with coefficients defines an ECC over a prime field. A node's closeness centrality (is also called closeness), which is determined by summing the shortest paths between it and another node which is present in the graph, is a measure of a node's centrality in a connected network. Because of this, the more important a node is in a network or linked graph, the closer it is to all other nodes. For an instance, some persons would be able to send a message to everyone else in the network very rapidly (few steps), but others may take several steps, if information needs to move across the network, therefore lower centrality score determines that a more central position or we can say important position.

Mathematically, closeness centrality is represented as,

$$C(i) = \frac{1}{\sum_j d(j,i)} \tag{5}$$

D(j, i) is a representation of the distance between the vertices j and i. Coming to its normalized form, where instead of sum the average shortest path lengths were considered, is represented as

$$C(i) = \frac{N}{\sum_j d(j,i)} \tag{6}$$

D. Eigenvector Centrality

In graph theory, a metric called eigenvector centrality is employed to measure a node's power inside a network. Based on the hypothesis that nodes with high scores are connected to contribute more to the node in question's score than it offers low-scoring nodes equal connections in relative rankings for every network node. This eigenvector centrality is determined by, doing a matrix calculation with an adjacent matrix to get the principal eigenvector. Not only this eigenvector centrality is important to find the influence in social networks but also this is used to rank web pages.

Consider a graph S; S:=(P, Q) which is of P vertices, A=($\mathbf{a}_{p,x}$) be the adjacency matrix, where ap,x=1. Vertex p is straight away linked x. Then relative centrality value for the vertex p is said as

$$\mathcal{X}_p = \frac{1}{\lambda}\sum_{x\in M(p)} \mathcal{X}_x = \frac{1}{\lambda}\sum_{t\in G} \mathbf{a}_{p,x}\mathcal{X}_x \tag{7}$$

λ is said to be constant and M(p) is the set of neighbours of vertex p.

E. Modularity

The measure of the structure of a network/graph is called "modularity". It assesses how well a network is divided into modules (which can also be referred to as groups, clusters or communities). Modularly high networks feature sparse connections among nodes in different clusters but dense connections among nodes within clusters. In optimization techniques for identifying community structure in networks, modularity is frequently employed. Modularity can recognize large communities, however, it has been demonstrated that it has a limit of resolution which cannot be used to detect smaller communities in the network.

Modularity Q is said to be $\frac{\sum \mathbf{A}_{ij} - \mathbf{k}_i\mathbf{k}_j}{2m}$ by all pair of i,j vertices of same group. Where Aij is the total number of edges in between i and j vertices (generally this value will be 0 or 1, but might have different values when networks have multiple edges). If edges are took place at random, the total number of vertices in-between I and j is given by $\frac{\mathbf{k}_i\mathbf{k}_j}{2m}$.

F. *Literature Survey*

Social Media Interfaces have become irresistible in today's workforce, sharing every kind of information, and that information doesn't necessarily be positive. These interfaces are being used in political strategies to spread the word among the people. As it is the fastest mode of passing out information. One such use case is Political Campaigns on Twitter. The purpose of this journal is to conduct an analysis of how close a particular presidential candidate is to a user [4].

We count the number of times a particular medium has been used The outlet cited various think tanks and political groups, Compare this to the times members of Congress cite the same group. All findings related to messages only Content; i.e., editorials, letters, etc. are excluded. Our findings demonstrate a clear leftist bias, with all news organizations receiving scores that were to the left of the typical member of Congress, with the exception of Fox News Special Report and the Washington Times. The New York Times and CBS Evening News got scores much to the left of the centre, supporting the assertions of conservative opponents. *PBS, CNN, and abc* were the media outlets with the most centrist viewpoints; USAToday was the most impartial print publication [2].

News sharing study was studied statistically and qualitatively to spot broad trends and elucidate more specific results. Users who share news, material, and networks were chosen as the three main study topics and were extensively examined. The core conclusion part uses the review's findings to offer a critical assessment of the state of the field and recommendations for future work [2]. The discussions in this literature review have been obtained by searching the CMMC and the ACM Digital Library databases. These discussions were seen between 2004 and 2014. For two reasons, we looked at this time period: The year 2004 was picked as the beginning point first because (a) Facebook was currently the biggest social media.

They proposed an empirical measure to quantify the dynamics and degree of 'bias' dynamics in mainstream and social media (henceforth referred to as news or blogs). The measurements are not prescriptive judgments but look at the attributes of the null model in the "unbiased" report above to look for bias. Data is based on RSS feeds aggregated by Open Congress. Open Congress is a nonprofit, nonpartisan public resource website that brings together official government data and the latest information on what's happening in Congress. We continuously monitor and collect the Open Congress RSS feeds of individual Congress members. This essay analyzes news and blog reports about the 11th Congress. [1]

The key difficulty for news-sharing research over the coming years will be keeping up with current and quickly evolving media advances, particularly those pertaining to mobile and visual communication. According to preliminary estimates, the messaging programmer WhatsApp has the potential to significantly increase traffic to news websites [5].

Data is taken from the news channels of fox news special, CBS evening news, new york times and many popular American news channels. The majority of the news outlets we analyze got points to one side of the normal individual from Congress. The paper did not offer a strategy for connecting this metric to the ideological positions of other political players. Their sample dataset contains this information and they excluded personal and in-person things like editorials, and letters to the editor [2].

This paper [1] develops a system-wide bias measure to quantify bias in mainstream and social media based on how often the media refers to members of the 11th US Congress. In addition to empirical measures, they also present a generative model to study how the global distribution of reference numbers per legislator in each media evolves over time. They plan to continue working along the lines of long-term tracking of propensity dynamics in two media, modelling of individual outlet biases, and use of content and multivariate analysis. We also gone through a case study about various types of Data Visualization techniques like word clouds, connective charts, Heatmaps and many more and also real world applications over Data Visualization. This case study focuses mainly on analyzing data visualization methods to examine how Facebook's micro-advertising affected the US election campaign [6].

For many years, the trustworthiness of the mainstream media and its dedication to "fair and balanced reporting" has been a topic of scholarly discussion. The 2012 presidential campaign saw the most significant resurrection of this interest. Based on the primary technique For many years, the trustworthiness of the mainstream media and its dedication to "fair and balanced reporting" has been a topic of scholarly discussion. The 2012 presidential campaign saw the most significant resurrection of this interest [3]. A meta-research of opinion surveys is conducted. According to statistics, the 2012 presidential election took place amid a period of American history that has been unquestionably marked by the people of the United States' mistrust of their elected government for more than a decade.

The authors argue that the scope of the results collected is severely limited and, as a result, most of the statistical underpinnings of the studies are dependent on the sources selected, thus the perceived perceptions this paper seeks to document. We admit that we cannot speak conclusively about the origin or extent of bias. While qualitative research has the potential to provide quantifiable insights into the research topic under consideration, is somewhat subjective in nature and is often restricted to examining and evaluating only quantitative studies in their academic framework [3].

III. Results & Discussions

A. Proposed Work

We used a simple concept that whichever text node with the given hashtag in the given network has more links must be having more out- degree and also must have more betweenness centrality. Therefore, the node with more betweenness centrality is the tweet posted by the candidate who is politically influenced on Twitter [4], [11].

Tools used for this analysis are Gephi and Rstudio. Gephi is a tool that helps in easy to make graphs, fast, can import data directly from CSV files and visualising interactively. There are many features to calculate subgraphs or metrics of the graph like in degree out-degree closeness centrality and many more other metrics which help us to study the networks and graphs easily.

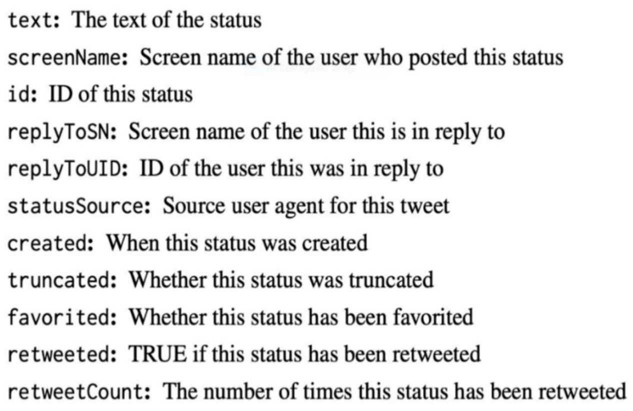
B. Requiring Dataset

The dataset required for this analysis is obtained from Twitter Database using the R Package ’twitteR’. It is an API for R Language. Our project is not based on any reference project and it is new. There are projects on media bias. But this is on political bias. We use the same package to convert the obtained data objects to Data Wire Frames.

```
#Install the Required Packages
> install.packages(”rio”)
> install.packages(”twitteR”)

#Requiring Libraries:
> library(twitteR)
> library(rio)

#Consumer key and Access Tokens:
CK <- ’Nt7yxxxxxxxxxxxxxx’
CS <- ’TV1oZoxxxxxxxxxxx’
AT <- ’706xxxxxxxxxxxxxxxxx’
ATS <- ’G6Axxxxxxxxxxxxxxx’

#Taking access from Twitter:
setup_twitter_oauth(CK,CS,AT,ATS)

#Searching Twitter
trump <- searchTwitter(
        ”#maga”,n = 5000,
        resultType = ”recent”)

joe <- searchTwitter(
        ”#joebiden”,n = 5000,
        resultType = ”recent”)

length(trump)
length(joe)

#twListToDf()
#converts tweets data file into a DF
trumpdf <- twListToDF(trump)
joepdf <- twListToDF(joe)
```

C. Pruning the Dataset

We define a few components to help us find relations and identify the frequency of users. We get the following fields from the API.



text: The text of the status
screenName: Screen name of the user who posted this status
id: ID of this status
replyToSN: Screen name of the user this is in reply to
replyToUID: ID of the user this was in reply to
statusSource: Source user agent for this tweet
created: When this status was created
truncated: Whether this status was truncated
favorited: Whether this status has been favorited
retweeted: TRUE if this status has been retweeted
retweetCount: The number of times this status has been retweeted

Fig. 1: Fields obtained in the dataset

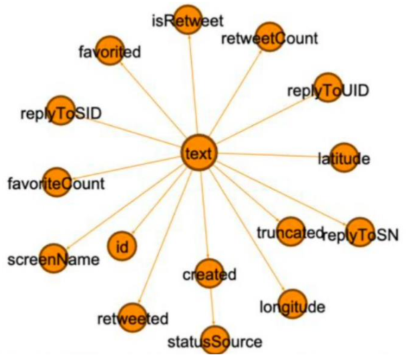


Fig. 2: Dataset at unit level

As we look few fields are useless and don’t show impact even if we remove those from the network. By default, we have some attributes populated with *NA* and *FALSE* values entirely. So, we can conclude these fields are not necessary. Some fields have a similar duplicate attribute (Name and ID), we can remove the ID fields from the network as it gives the same result without presence. So by pruning, we are left with ’Five’ attributes that are significant for our network formation and measures.

```
#Removing the Insignificant fields from DF.

trump <- trumpdf[,c(1,4,11,12,13)]
joe <- joepdf[,c(1,4,11,12,13)]

#Viewing and Exporting the DF
```

```
View(votefortrumpdf)

export(trump, ”trump.csv”)
export(joep, ”joe.csv”)
```

The fields that are more significant are *text, screenName, replyToSN, isRetweet, retweetCount*.

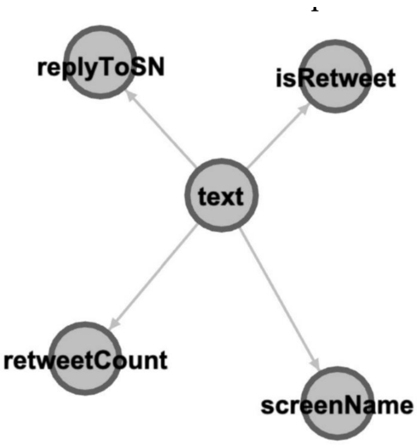


Fig. 3: Dataset after pruning.

D. Visualisation

We now have a pruned dataset to visualize and conduct the centrality measures. We use the software Gephi for visualisation as it is user-friendly and also accurate enough.

1) We first need to import the dataset into the software.
2) Import it as an Adjacency list because the dataset we are using will already have both nodes and edges included.
3) Time Stamps are automatically selected w.r.t the type of dataset.
4) Now, the network is formed and visible to us. To make it clear and understandable, use a layout called ’Force Atlas’ and change the value of ’Repulsion Strength’ to 10,000. To view, the node labels click on **T**
5) After getting the desirable layout, Conduct the centrality measures Network Diameter (Includes Betweenness and Closeness), Eigen Vector Centrality, Average Degree and Modularity.
6) Here we can visualise the network using the centrality measures by utilizing the two options. It adds colours to the nodes and also sizes them based on the betweenness centrality (Optional Step).

   Fig. 4 and 5 show the visualisation done on the dataset after pruning.

E. Results

The following are the visualisations of the various centrality measures of the networks along with values of few basic graph metrics calculated against the network inside the tool.

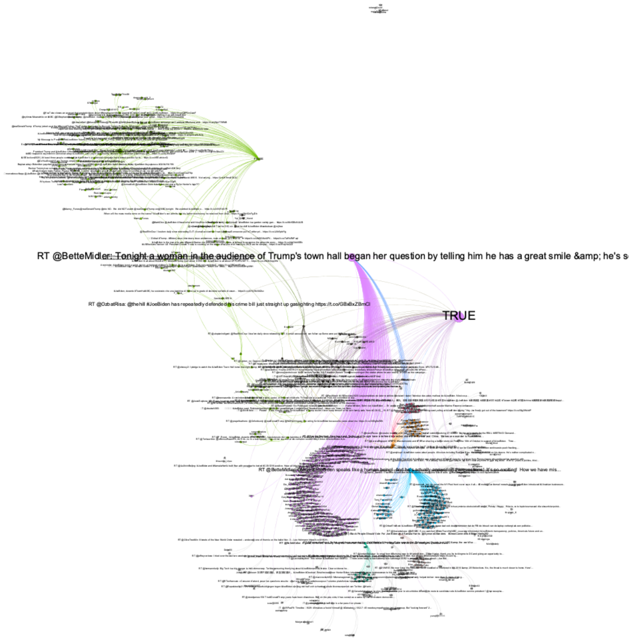

RT @BetteMidler: Tonight a woman in the audience of Trump's town hall began her question by telling him he has a great smile &amp; he's s
TRUE

Fig. 4: Joe’s Network.

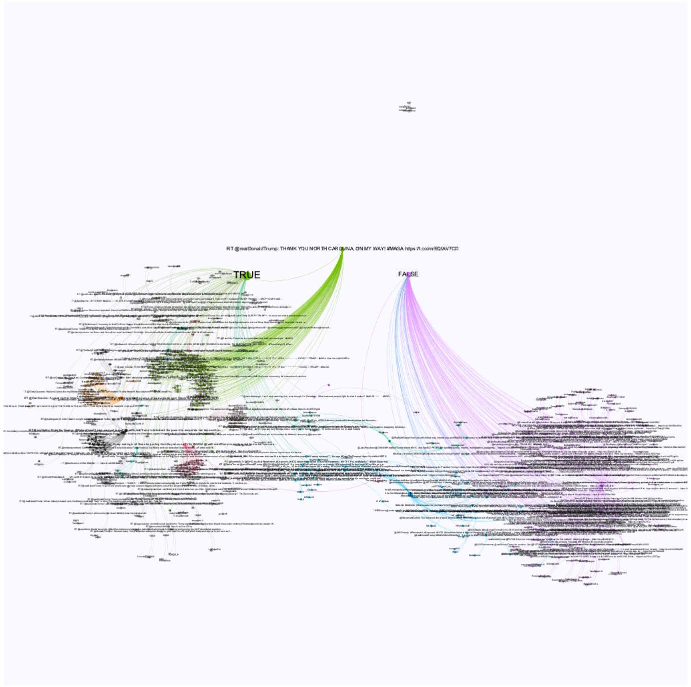
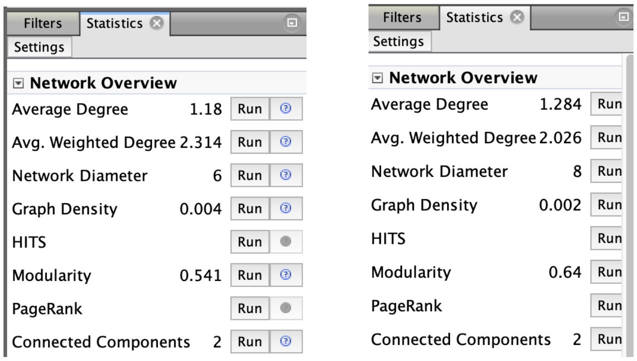

RT @realDonaldTrump: THANK YOU NORTH CAROLINA, ON MY WAY! #MAGA
TRUE
FALSE

Fig. 5: Trump’s Network.



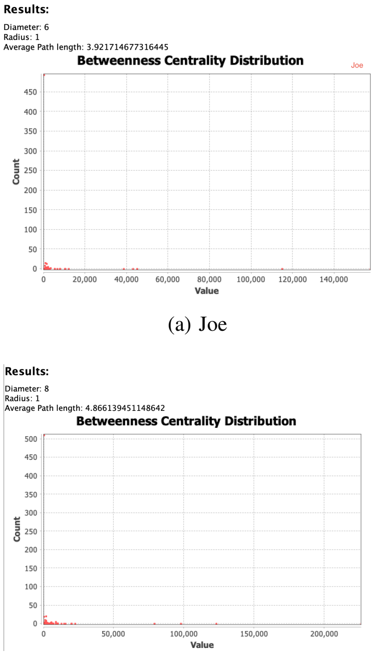
Network Overview
Average Degree 1.18 Run
Avg. Weighted Degree 2.314 Run
Network Diameter 6 Run
Graph Density 0.004 Run
HITS Run
Modularity 0.541 Run
PageRank Run
Connected Components 2 Run

(a) Joe

Network Overview
Average Degree 1.284 Run
Avg. Weighted Degree 2.026 Run
Network Diameter 8 Run
Graph Density 0.002 Run
HITS Run
Modularity 0.64 Run
PageRank Run
Connected Components 2 Run

(b) Trump

Fig. 6: Various Graph Measures


Results:
Diameter: 6
Radius: 1
Average Path length: 3.921714677316445

(a) Joe

Results:
Diameter: 8
Radius: 1
Average Path length: 4.866139451148642

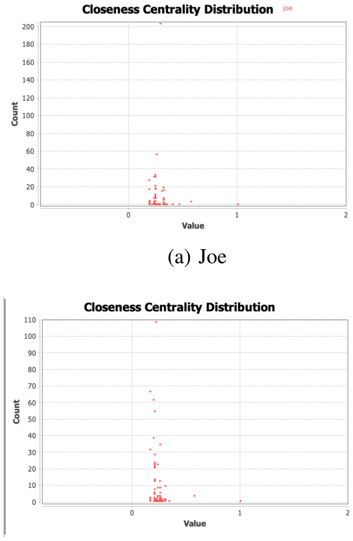
(b) Trump

Fig. 7: Betweeness Centrality Measures

After visualizing and analyzing we get the influential tweet with the hashtag #joebiden has a betweenness centrality of (1,14,726), out-degree of (221) and eigenvector centrality of (1). On the other hand, the influential tweet with #maga has a betweenness centrality of (78,672), out-degree of (116) and eigenvector centrality of (0.4).

TABLE I: Comparision Results

Measure	Joe Biden	Donald Trump
Out-Degree	221	116
Betweenness	1,14,726	78,672
Eigen-Vector	1	0.4
Modularity	37	50
Closeness	0.4	0.3



(a) Joe

(b) Trump

Fig. 8: Closeness Centrality Measures

IV. Conclusion

We can say that Network Analysis helps in finding out the most influential candidate on various social networks. The results of this project are the analysis of political bias on social media to identify the key aspects that are involved in an influential status. We can analyze this using the software Gephi for plotting the graph and calculating the betweenness centrality and other centrality measures. Even though more no.of retweets are being done to the tweets with #maga, We can conclude from the above results that Joe Biden was a more influential candidate than Donald Trump on Twitter.

References

[1] LIN, Y.R., BAGROW, J. P., and LAZER, D. 2011. More voices than ever? quantifying media bias in networks, In ICWSM, The AAAI Press.

[2] GROSECLOSE, T. AND MILYO, J. 2005. A measure of media bias. The Quarterly J. of Economics 120, 4, 1191–1237.

[3] Quackenbush D (2013) Public perceptions of media bias: A meta-analysis of American media outlets during the 2012 presidential election. The Elon Journal of Undergraduate Research in Communications 4: pp.1-6.

[4] Zinoviev, Dmitry. (2018). Network Analysis of the 2016 Presidential Campaign Tweets.

[5] Kümpel Anna, Karnowski Veronika, Keyling Till. (2015). News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks. Social Media Society. 1. 10.1177/2056305115610141.

[6] Kochhar.D, Meenakshi.S.P, Dubey.S.(2020). Applications of Visualization Techniques. In: Anouncia.S, Gohel.H, Viramuthu.S(eds) Data Visualization. Springer, Singapore. 10.1007/978-981-15-282-6-6.

[7] Mochamad Suyundi, Abdul Talib Bon. Determinig the Size of Centrality in Social Networks. 11th Annual International Conference on Industrial Engineering and Operations Management Singapore, March 7-11, 2021.

[8] Karikalan Nagarajan* , Malaisamy Muniyandi, Bharathidasan Palani and Senthil Sellappan. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. 10.1186/s12874-020-01119-3

[9] Ning Li, Qian Huang, Xiaoyu Ge, Miao He, Shuqin Cui, Penglin Huang, Shuairan Li, and Sai-Fu Fung. A Review of the Research Progress of Social Network Structure. Hindawi, Complexity, Volume 2021, Article ID 6692210. https://doi.org/10.1155/2021/6692210.

[10] Poonam Sharma. Centrality Measures in Social Networking: Study and Analysis Using NetDraw 2.138 in UCINET6. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 3 Issue 4, April 2014.

[11] Peng, S., Yang, A., Cao, L., Yu, S., & Xie, D. (2017). Social influence modeling using information theory in mobile social networks. Information Sciences, 379, 146–159

[12] Min, B., Liljeros, F., & Makse, H. A. (2015). Finding influential spreaders from human activity beyond network location. PLoS ONE, 10, e0136831

[13] Senthil Murugan, N., & Usha Devi, G. (2018). Detecting streaming of Twitter spam using hybrid method. Wireless Personal Communication, 103, 1353–1374. https://doi.org/10.1007/ s11277-018-5513-z