Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PUBLIC OPINION, CLUSTERS, AND STATISTICS OF COVID-19 VACCINATION: A TWITTER ANALYSIS

GHADA AMOUDI[1], ZELAL SHEARAH[2]

1, 2 Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

E-mail: [1]gaamoudi@kau.edu.sa, [2] zmahmoudshearah@stu.kau.edu.sa

ABSTRACT

The coronavirus disease (COVID-19) pandemic has drastically affected the entire world. Vaccinations have been developed to contain the spread of the virus and help humanity recover from the pandemic. However, people are often reluctant to adopt new medical interventions, and COVID-19 vaccines are no exception. Social media platforms like Twitter are flooded with discussions, opinions, and rumors about vaccines. Numerous people access news regarding COVID-19 vaccines on Twitter—rather than through official media channels—which offers a wealth of comments about news and medical breakthroughs. However, as people are panicked due to this alarming situation, they tend to share any information they receive without checking its credibility, creating more panic. Analyzing interaction between social media users provides insight into the spreading pattern of news, people's opinions towards a particular issue, who the influencers are, and how they influence others, among other findings. This study investigated opinions about COVID-19 vaccines by applying sentiment analysis and detecting communities on Twitter. We constructed an interaction network of discussions related to COVID-19 vaccines and applied social network analysis methods to find communities and nodes with high centrality measures. Next, we analyzed how these nodes affect the overall community opinion. Two main communities were detected, with the larger community displaying a higher positive sentiment ratio than the smaller one. Furthermore, the polarity of the high centrality nodes in each community was close to the average polarity of the community as a whole. These findings highlight the potency of the node's position in terms of centrality measures. In conclusion, analyzing discussion networks should not be overlooked when public health is concerned, as influencers are not necessarily those with high numbers of followers but rather those with high centrality measures within the interaction network regarding the topic being discussed.

Keywords: *Community Detection, Girvan–Newman, Sentiment Analysis, Centrality, Betweenness*

1. INTRODUCTION

The world was severely affected by the deadly coronavirus disease (COVID-19), which spread rapidly in a short period, causing chaos and fear across countries. According to World Health Organization (1), more than 248 million COVID-19 cases and five million deaths were recorded worldwide between 2019 and 2021. Therefore, significant research and global health institutions have been directed toward producing vaccines to stop the proliferation of COVID-19 and reduce the effects of the disease. Medical research has successfully produced vaccines, providing a significant breakthrough in managing the spread of the virus (2). However, the COVID-19 vaccines are accepted by some people and rejected by others. A large amount of data is available online as people discuss their opinions about vaccines on various social media platforms. Twitter, a popular microblogging site, has played an essential role in public discussions about COVID-19 vaccines (2). The common assumption that influencers or social media accounts with high numbers of followers have a significant impact on public opinion is not always valid. A social network such as Twitter consists of "nodes" (i.e., users) and "edges," which can be represented by various connections or relationships, such as following, retweeting, or replying. The underlying network structure, that is, how nodes and edges are connected, plays a vital role in determining influencing nodes. However, the network structure is merely a part of the big picture; usually, connectedness involves two related elements, structure and behavior; that is, the action of a node has implicit outcomes on the whole network (3). Accordingly, to analyze a social network and determine how information flows within a network, a careful examination of its structure and dynamics should be conducted.

Governments have made great efforts to create healthcare campaigns to encourage people to get vaccinated against COVID-19; nevertheless, some individuals oppose these efforts by spreading rumors and misleading information about the risks related to vaccines. Such behavior, especially on social media platforms, could have a substantial adverse effect, hindering the government's efforts to curb the spread of COVID-19. Social Network Analysis (SNA) methods provide measures to describe, analyze, and process social networks effectively. Centrality measures, for instance, can be used to determine powerful nodes through many different means, not only by counting the number of followers. Thus, employing SNA in social media conversations would reveal hidden communication patterns that could help stakeholders mitigate the negative impact created by this impetus of discussions. For example, the betweenness centrality of a node determines how important it is in transferring knowledge among different communities. In contrast, closeness centrality measures the extent to which a node is close to every other node in the network (3). Community detection techniques, such as the Girvan–Newman algorithm (4), can dissect a network and divide it into many smaller networks, based on the betweenness measure of the edges. Community detection has numerous applications, such as finding terrorist communities (5) and discovering the hidden community structure of countries affected by COVID-19 (6). Detecting communities can be achieved by numerous methods, including machine learning, deep learning, and graph-based methods. In this study, we aim to explore various centrality measures and how they relate to communities. Thus, graph-based methods were adopted. This study presents an analysis of a sample of tweets discussing COVID-19 vaccines. First, the sentiment of each tweet is determined, and a network is constructed based on the replies among users. Once the network is created, centrality measures are derived, and community detection is conducted. Finally, the average sentiment of each community is compared with that of high centrality nodes within the community. Formally, the research question for this study can be framed as follows:

1. How can we extract communities formed around COVID-19 vaccine discussions?
2. How do high centrality nodes in a community influence other users' opinions within the community?

To answer the first research question, Girvan-Newman community detection algorithm was applied. While the second question was answered by finding sentiment for each user and community, calculating centrality measures to determine influencers, and finally contrasting average community sentiment with high centrality nodes sentiment. Addressing these research questions provides insights into public opinion regarding COVID-19 vaccines; more importantly, it contributes to the knowledge of the interaction network structure and how it affects overall public opinion. In other words, dissecting social media users into groups based on interaction (discussion) in a given scenario reveals how people form opinions and attitudes towards a particular issue. Analyzing users' sentiment during a controversial topic, such as the COVID-19 vaccine, is a standard procedure. In this work, however, we examine how the average community sentiment reflects the polarity of the influencers within the community. Thus, the potential contributions of this study can be summarized as follows:

1. A dataset of recent tweets (approximately 5,800) was collected using relevant COVID-19 vaccine keywords.
2. Sentiment analysis of tweets, communities of users based on their interactions, and how communities' average sentiment compares with the sentiment of high centrality nodes within each community.

The remainder of the paper is organized as follows: in the following section, a literature review is presented. Next, the methodology and dataset are described, followed by the results and discussion. Finally, the conclusions and avenues for future work are discussed.

2. LITERATURE REVIEW

News and information about the COVID-19 pandemic have dominated media outlets and social media unprecedentedly, which has contributed to shaping public opinion and government health measures (7). Social media, such as Twitter, provides information and opinions on global issues such as COVID-19 vaccines (8). Studies have used different techniques to determine how people have discussed COVID-19 vaccines on social networks and whether social media use played a role in accepting or rejecting the vaccines. We present the efforts done to analyze people's attitudes towards the COVID-19 vaccine; next, we focus on the technical perspective of the literature.

In (9), Twitter polls were utilized to investigate people's desire to be vaccinated against COVID-19 and to understand popular perceptions about vaccine safety. The results showed that while most people were uncertain about the safety of COVID-19 vaccines, they were willing to be vaccinated. It also confirmed that social media greatly influence public opinion and views about health issues. The acceptance rate of vaccines among the general population in different regions of the world was investigated to compare acceptance rates during the pre-and post-approval period (7). This study also identified areas with low vaccine acceptance rates. This, in turn, can help government agencies and public healthcare officials address their concerns. Additionally, the study suggested that confidence in COVID-19 vaccination can be consolidated among the general population by disseminating messages about the protection and efficacy of the available COVID-19 vaccines through trustworthy networks. It found that the highest acceptance rate was in the western Pacific region (67.85%) and the lowest in the African region (39.51%). Authors in (11) conducted a study in the United States to analyze the spatiotemporal patterns of general feelings and emotions related to COVID-19 vaccines among Twitter users to reveal potential drivers of changing sentiment and emotion over time. The results showed that real-time analysis of social media data benefits health authorities by enabling them to monitor the public's opinions and attitudes toward information related to the COVID-19 vaccines by region, thereby helping them address the concerns of vaccine skeptics and enhancing the confidence of individuals within an area or a specific community. In (12), the authors conducted a study in Canada to uncover the reasons underlying the reluctance to receive COVID-19 vaccines. The results revealed that safety, lack of sufficient information about the vaccine, rumors based on conspiracy theories, and unclear legal responsibilities were among the reasons that caused uncertainty about COVID-19 vaccines. Similarly, public opinion regarding accepting COVID-19 vaccines in the United States was investigated (13). They analyzed Twitter posts and found that influencers significantly impacted public opinion.

The effect of vaccination discussions on people's acceptance of vaccines was examined (14). They also investigated how networks of skeptics and supporters were shaped differently. The results showed that anti-vaccine community groups demonstrated greater consensus than pro-vaccine community groups. They also found that positive sentiments are more dominant in public discussions associated with significant events; this indicates a higher acceptance of this vaccine than in previous ones. Authors in (15) studied the public's interests, views, and feelings about COVID-19 vaccines, as discussed on Twitter, and found that public interest revolved around three main topics: information about vaccine reservations, COVID-19 infection and mutation, and vaccination status in other countries.

The authors in (10) applied content and sentiment analysis to COVID-19 vaccine-related tweets by analyzing 2.4 million relevant tweets in 2020. They also analyzed user accounts to reveal public perceptions of the COVID-19 vaccine. The results showed that most people are optimistic and confident about the vaccines. Public sentiment on Twitter during the COVID-19 pandemic was analyzed using two datasets—Indian and global (16). The authors created a model using Bi-directional Encoding Representation for a Transformer (BERT). They automatically labeled the tweets using Valence Aware Dictionary and sEntiment Reasoner (VADER) (17) and TextBlob (18), a Python library for common text processing tasks, including sentiment analysis. The study performed a multiclass classification using seven different polarity measures—ranging from strongly positive to strongly negative—using the scores from VADER. The result of the BERT-based classifier showed 94% accuracy, while the sentiment results showed that the global tweets were more neutral than the Indian tweets and that more strongly negative and strongly positive tweets appeared in the Indian dataset. Deep learning methods for sentiment analysis were compared in (19). The authors proposed a text vectorized neural network model to classify opinions in Twitter posts regarding COVID-19 vaccines. The model achieved an accuracy score of 0.81, which outperforms long short-term memory (LSTM) and bidirectional long short-term Memory (BiLSTM) models. Another comprehensive comparison was conducted in (20) to measure opinions and hesitancy regarding COVID-19 vaccines. The study analyzed Twitter posts using three different sentiment analysis tools, VADER, TextBlob, and Azure Machine Learning, with five learning algorithms, three word vectorization methods, and three different stemming approaches; a total of 42 experiments were conducted. The best performing model was TextBlob with TF-IDF and linear support vector machine using stemming and lemmatization; the accuracy achieved was 0.967. The study also stated that public hesitancy decreased gradually over time as more people took the vaccine and more awareness spread among communities.

Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Bhatnagar and Choubey (21) applied community detection and sentiment analysis to Twitter data to cluster users based on their discussions on predefined topics. The network was constructed based on hashtags, which denotes a node connected to another node if the tweet posted contained the same hashtag. They applied a fluid community detection algorithm (22) and supervised machine learning for sentiment analysis. The results revealed that different communities share the overall sentiment for each topic, with a largely negative polarity for COVID-19-related issues. Chaudhary and Singh (6) analyzed a COVID-19 dataset of confirmed cases worldwide from Johns Hopkins University's official website. The study applied unsupervised machine learning to detect communities and principal component analysis (PCA) to reduce dimensionality, finding that community detection improved after using PCA.

The literature review shows that many attempts have been conducted to make sense of COVID-19 data, ranging from medical data to social media discussions. Several studies have analyzed Twitter data to mine people's opinions regarding vaccines and other topics using sentiment analysis and text mining techniques. Others have used Twitter polls, while a few have applied community detection methods. While these studies provide informative results regarding peoples' acceptance of vaccines and other COVID-19 concerns, no study has approached the vaccine acceptance issue in conjunction with community detection. Although authors in (21) applied community detection and sentiment analysis to Twitter data; however, their study included various topics and was not focused on the COVID-19 vaccine.

Moreover, researchers have investigated public opinion regarding the vaccine using mainly machine learning approaches; our approach combines graph-based social network analysis and machine learning-based sentiment analysis. Therefore, this study analyzes Twitter posts discussing COVID-19 vaccines to identify people's opinions and determine how the underlying network structure influences their beliefs. The study applies a community detection algorithm to find clusters of users in a network based on their interactions. Combining community detection with sentiment analysis techniques provides knowledge regarding the public acceptance of COVID-19 vaccines with respect to the community. This knowledge will provide insights into the mutual influence of community members, which could, consequently, help authorities identify the real influencers within an interaction network discussing a particular topic.

3. RESEARCH METHODOLOGY

This section describes the methodology and steps followed in this study: data collection, data pre-processing, sentiment analysis, and community detection methodology. Figure 1 illustrates the flow of the research methodology.

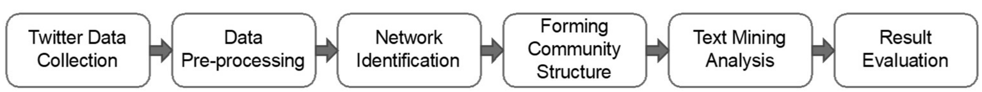

Figure 1 Research methodology workflow

3.1 Data Collection

Twitter is considered one of the world's major social media platforms, with 187 million daily active users in the third quarter of 2020 (23). Netlytic service (24) was used to collect COVID-19 vaccination-related tweets. Netlytic is a cloud-based social network analyzer that helps uncover hidden patterns in social networks from online conversations from several sources, including Twitter, Facebook, YouTube, text files, RSS feeds, and cloud storage (24).

The data used in this study comprised 5,816 tweets on COVID-19 vaccination posted between November 11, 2021, and December 11, 2021. These vaccine tweets were collected from all over the world using search terms such as coronavirus, vaccine, Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V.

3.2 Data Pre-Processing

To determine the sentiment of each tweet, the VADER text mining package (17) was employed. Text mining tools, including VADER, use only English words and regard each word as a feature in a feature vector. Thus, the feature vector could include thousands of words, complicating the classification task. A common practice to reduce the dimensionality of this vector is to pre-process the words input into the classification model by removing words and symbols that do not significantly contribute to the sentiment of the tweet. Tweets are generally "noisy," including punctuations, emojis, URLs, stop-words, misspelled words, and numbers. Pre-processing aims to clean the text and prepare it for the classification model. One way of addressing emojis, URLs, and numbers is to replace them with the words "EMOJI," "URL," and "NUMBER" (25). Another approach is to eliminate them (2). The steps adopted for pre-processing data for this study are as follows:

1. Converting the string of tweets to lowercase: This is a common practice when applying any text mining task to reduce dimensionality, as the machine considers the same word with different letter cases as two different words.
2. Removing emojis: While emojis would add to the meaning and especially to the sentiment of the sentence, they were omitted in this study because the tool employed for sentiment analysis did not handle emojis. To keep emojis in the feature vector, a few more steps may be added to the pre-processing algorithm, which could be an avenue for future research.
3. Removing meaningless characters, stop-words, and punctuation, as they have no semantic content: Table 1 provides examples of meaningless symbols, punctuation, and stop-words.
4. Removing URLs (http://www.situs.com), email (name@situs.com), hashtag symbols (#), and symbols (@).

Table 1 *Examples of meaningless characters, punctuation, and stop-words*

Character	Punctuation	Stop-words
**	,	I
SSS	()	me
=	;	Myself
$	-	during
&	~	before

3.3 Network Building and Community Detection

In this step, an interaction network graph between users in the sample dataset was constructed. In this context, users are nodes, and edges represent the "reply to" relation; thus, the graph was created by adding an edge between two nodes if a user replied to another. Consequently, we constructed an undirected network, as we only needed to capture the interaction among users, regardless of the direction in which the interaction took place. Subsequently, community detection was performed, and nodes with high closeness centrality were identified.

Community detection is the process of discovering interconnected groups or clusters in a network. The goal of community detection is to find groups of nodes that are, in some sense, more similar to each other than to other nodes outside the community. To find the communities in our network dataset, we used the Girvan–Newman algorithm (4), which finds the edge with the highest betweenness centrality and removes it. This edge is considered a bridge that connects two or more clusters of nodes (3). The resulting clusters function as communities because the bridge that connects the cluster with other clusters is removed. The Girvan–Newman algorithm consists of three main steps. First, the algorithm computes the betweenness centrality for each edge and eliminates the one with the highest score. Second, the algorithm recalculates the betweenness for every edge and removes the one with the highest value. Third, this process is repeated until no more edges are left or every edge has the same betweenness centrality (3).

Centrality measures calculate the connectivity of each node within a network and reflect the importance of nodes (26). The most common centrality measures are degree, closeness, and betweenness centrality. Degree centrality depicts the number of direct nodes connected to a specific node, while closeness centrality refers to the average shortest distances to all other nodes in the network. Betweenness centrality measures the extent to which a node lies in the shortest path of other nodes (26). Unlike node betweenness, edge betweenness represents the total amount of flow the edge carries between all pairs of nodes using this edge (3). In this study, degree and closeness centrality were calculated for each node in each community, and the top five nodes were highlighted for each. These two measures were selected because they reflect the importance of nodes within a community rather than

Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

betweenness centrality, representing the importance between communities.

3.4 Network Visualization

To visualize the previous network, we used the Gephi software (27), a free network analysis tool that can be used to visualize large network graphs. It employs a 3D rendering engine for efficient graph visualization and provides other network calculation metrics, such as centrality measures. The data network on Gephi before any change or analysis is shown in Figure 2.

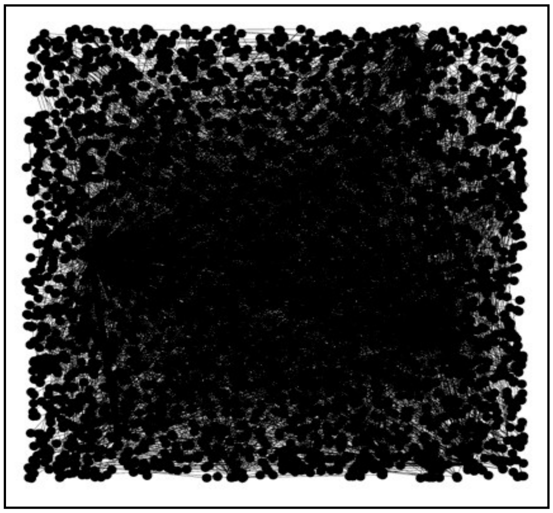

Figure 2 Network visualization by Gephi

Figure 3 depicts the network after applying the ForceAtlas layout, which is available in Gephi in the drawing options. ForceAtlas is a popular layout for presenting social networks more naturally and efficiently and reveals how edges connect the nodes while the nodes repel each other. The ForceAtlas technique allows a better visual understanding of the structure by converting the structural resemblance into a visual similarity and promotes a grasp of the network's important features. Active nodes within a community have more relations than with the outside and share high-density connections.

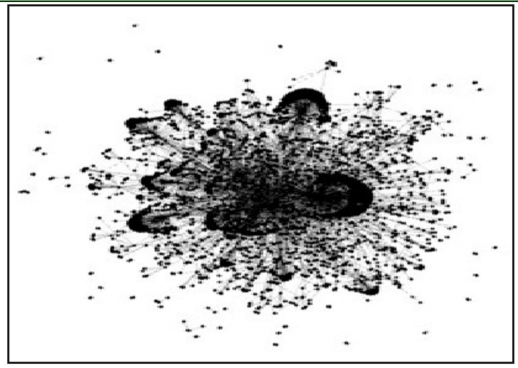

Figure 3 Network visualization by Gephi after applying the ForceAtlas layout

3.5 Sentiment Analysis

The data were primarily collected for sentiment analysis regarding vaccinations against COVID-19. Therefore, sentiment analysis was conducted to identify the polarity of each tweet to unveil the opinions expressed in the tweets. Each tweet was labeled as negative, neutral, or positive. To accomplish this, we used the VADER text mining tool, which contains, among other text mining tasks, a lexicon and rule-based component for sentiment analysis tailored explicitly for short social media posts (17).

4. RESULTS

After collecting data from Twitter and building a network, we used the Girvan–Newman algorithm to detect communities within the network. The algorithm works iteratively, eliminating edges with the highest edge betweenness and removing edges from the graph in a specified manner. This results in segmenting the network graph into smaller pieces or communities. Figure 4 illustrates the original network after applying the Girvan–Newman algorithm and setting different colors for each community.

Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

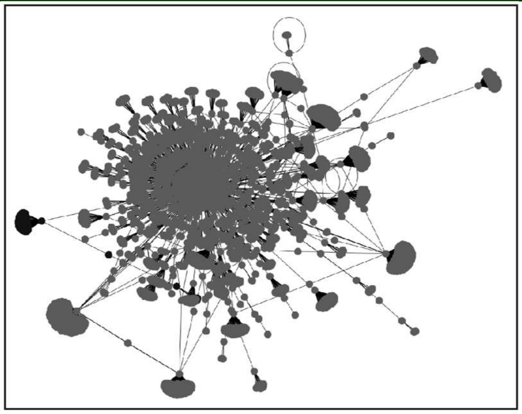

Figure 4 The original network

After applying the Girvan–Newman algorithm, the network was divided into two communities according to the results and partitioning. Figures. 5 and 6 depict the first and second communities, respectively. The first community contains 4,599 nodes and 4,786 edges, while the second has 100 nodes and 99 edges.

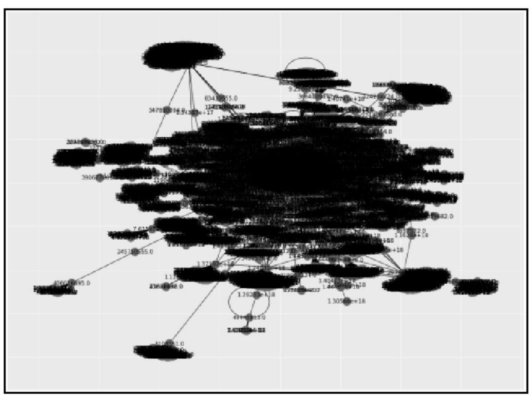

Figure 5 First community network

Sentiment analysis was conducted on three networks: the original network, the first community network, and the second community network. The sentiment analysis reveals that the words that appeared most frequently were "COVID-19," "vaccination," "COVID," "people," and "misinformation." Figure 7 illustrates the average sentiment across the tweets on the original network. Figures 8 and 9 show the average sentiment across tweets in the first and second communities.

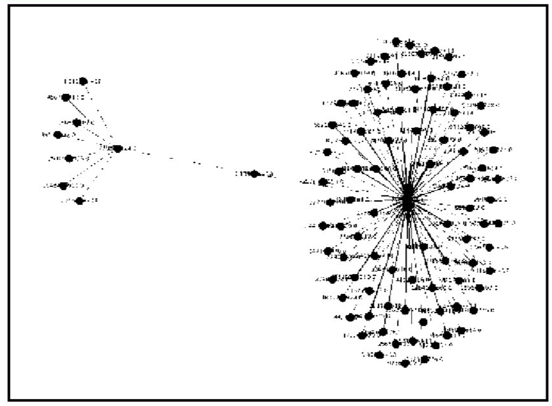

Figure 6 Second community network

The sentiment analysis revealed that in both communities, most tweets were neutral. Furthermore, it shows a similar ratio of neutral tweets in both communities; however, the second community contained more positive than negative tweets regarding the acceptance of vaccination. The first community showed similar positive and negative tweet counts.

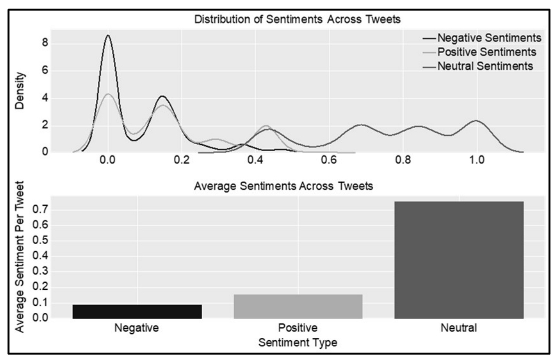

Figure 7 Original network sentiment analysis

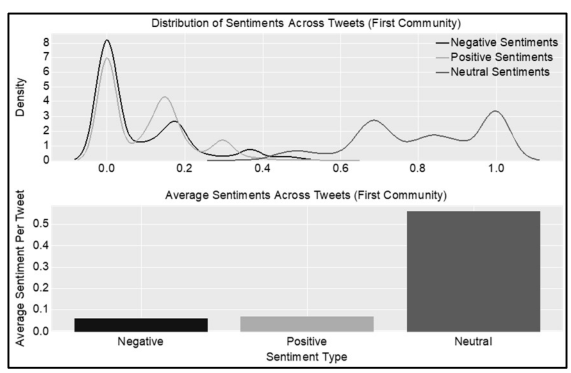

Figure 8 First community sentiment analysis

Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

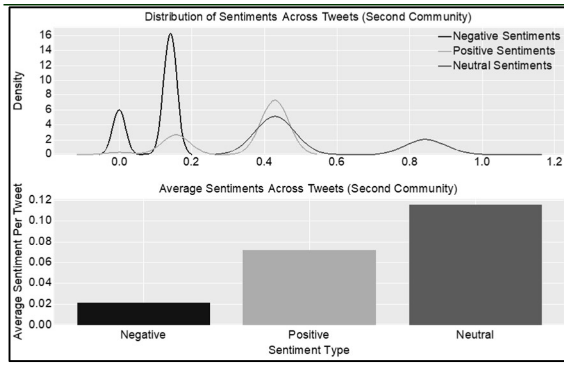

Figure 9 Second community sentiment analysis

Next, the top five high-degree centrality nodes for the first and second communities were identified separately. This step is essential to contrast the sentiment of those nodes with the average sentiment of the community in which they reside. Table 2 illustrates the sentiments for the top five high-degree centrality nodes in the first community; users' identification numbers were replaced by proxies for privacy purposes: for instance, C1.1 represents the node with ID 1 in community 1.

As presented in Table 2, the high degree centrality nodes have an average positive polarity of 0.17, an average neutral polarity of 0.77, and average negative polarity of 0.05. By comparing the average sentiments presented in Figure 8 and those of the high degree centrality nodes, we observe that the average sentiments of the community, to some extent, reflect the sentiments of the high degree centrality nodes; that is, the average neutral sentiments across all tweets and of the top centrality nodes (0.77) were the largest. The small positive and negative sentiment averages of the high centrality nodes—0.17 and 0.05, respectively— are also reflected in the averages within the community. However, the average positive and negative sentiments of the high centrality nodes are not equal as in the community. The closeness centrality was also obtained for both communities, and the top five for the first community are shown in Table 3. Nodes 1 and 3 in community 1 are the top two in closeness centrality; these two nodes appeared in both top lists and have a positive-neutral sentiment.

***Table 2** Sentiments for the top five high-degree centrality nodes in the first community*

User ID	Degree Centrality	Positive Sentiment	Neutral Sentiment	Negative Sentiment	Overall Sentiment
C1.1	0.0809	0.361001	0.549001	0.090001	Neutral–Positive
C1.2	0.0504	0.185001	0.815001	0.000001	Neutral
C1.3	0.0451	0.200001	0.698001	0.102001	Neutral–Positive
C1.4	0.0285	0.113001	0.887001	0.000001	Neutral
C1.5	0.0188	0.000001	0.912001	0.088001	Neutral–Negative
Average		0.17	0.77	0.05	

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3 Top Five High Closeness Centrality Nodes In The First Community

ID	Closeness Centrality
C1.1	0.3733
C1.3	0.3692
C1.6	0.3648
C1.7	0.3588
C1.8	0.3583

An example of a positive tweet from the top centrality nodes in the first community posted by node C1.3 is: "470,000 lives among people aged 60+ have been saved by #COVID19 vaccines since their roll-out in 33 WHO European Region countries." An example of a negative comment by user C1.6 is, "Over the past week, three children have died following their vaccination with the Pfizer Covid-19 vaccine in Bac Giang, Hanoi, and Binh Phuoc. The cause of death was determined as an 'overreaction to the vaccine.' A neutral comment by user C1.1 is "Lots of misinformation about vaccination doing the rounds again, with supporters of the ultra-libertarian Great Barrington Declaration out in force. Thus, here's a handy infographic from @britsocimm to share around. More information here:https://t.co/RB36ZdTWgkhttps://t.co/sl0vOO Khqy."

Table 4 illustrates the sentiments for the top five high-degree centrality nodes in the second community, where the sentiments are primarily neutral and then positive. Table 5 demonstrates the closeness centrality in the second community.

Table 4 Sentiments For The Top Five High-Degree Centrality Nodes In The Second Community

User ID	Degree Centrality	Positive Sentiment	Neutral Sentiment	Negative Sentiment	Overall Sentiment
C2.1	0.9191	0.333001	0.667001	0.000001	Neutral–Positive
C2.2	0.0202	0.185001	0.650001	0.650001	Neutral–Negative
C2.3	0.0101	0.300001	0.700001	0.000001	Neutral–Positive
C2.4	0.0101	0.300001	0.700001	0.000001	Neutral–Positive
C2.5	0.0101	0.300001	0.700001	0.000001	Neutral–Positive
Top Nodes Average		0.28	0.68	0.13	

Table 5 reveals that the top high closeness centrality nodes are the same as the top five high degree centrality nodes. This correspondence could be due to the community's relatively smaller size. Moreover, there was more consensus in the smaller community than in the larger community. The high degree centrality nodes have an average neutral polarity of 0.68, which is the largest among the three, and positive polarity of 0.28, which is larger than the average negative polarity (0.13). By contrasting these ratios with the distribution shown in Figure 9, we observe that the average neutral sentiment is the largest within the community, and the average positive sentiment is larger (approximately 0.075) than the average negative sentiment (0.02).

5. DISCUSSION

As applied research, we employed a community detection algorithm and an over-the-shelf tool for polarity classification; thus, the result interpretation criteria lie in the information obtained, the

Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

communities, the polarity scores, and the centrality measures, similar to (14) and (21). To answer the first research question, a social network was constructed from the interactions (replies) between Twitter participants discussing the COVID-19 vaccine. Community detection was applied to this interaction network to find clusters of users. Two communities were identified—a large community with 4,599 nodes and a smaller one with 100 nodes. This finding supports the observation that a social network can only have one giant component (3). Next, we computed the polarity for each post, the average polarity for each community, and the average polarity of the top five high-degree centrality nodes.

Table 5 Top five high closeness centrality nodes in the second community

ID	Closeness Centrality
C2.1	0.8685
C2.2	0.5052
C2.3	0.46610
C2.4	0.46610
C2.5	0.46610

A high centrality node within a community represents an active actor in the discussion and is not necessarily an influencer or a celebrity. However, these actors significantly influence the opinions of others because of their high centrality measures, which cannot be detected merely by viewing the number of friends and followers. This study revealed that the sentiment averages of the top five high-degree centrality nodes in each community are similar to the community in which they reside. The top five high closeness centrality nodes were obtained from both communities. In the large community, the results showed that two of these nodes are also in the top five high-degree centrality nodes, which emphasizes the importance of these two nodes in reaching the other nodes in fewer steps. In the smaller community, the top high closeness centrality nodes are the same as the top five high degree centrality nodes, confirming their status within the community. Thus, there is a resemblance between the high centrality nodes' sentiment and the community's average sentiment, which answers the second research question.

Several studies have investigated people's attitudes and sentiments toward COVID-19 vaccines. Some have analyzed sentiment within specific communities (7), (11), and (16), while others have not specified a particular region (9) and (10). In this study, no specific area was determined; however, the keywords used to find tweets were in English, which may have restricted the findings to English-speaking countries. Moreover, there is no consensus in previous research regarding public sentiment about COVID-19 vaccines, as some studies have revealed a mostly negative sentiment (21), while others have found it to be primarily positive (14). Our results agree with (14), as most tweets are neutral and positive.

Among the reviewed literature, some works focus on sentiment and clusters from a technical perspective, as in (21), (14), and (6). Methods used in these studies to detect communities fall into two categories: graph-based and machine learning-based. Chaudhary and Singh (6) applied a machine learning-based approach, whereas the other two and our study used graph-based methods. Nevertheless, the results of these studies confirm that communities usually share common views, supporting our findings.

This study focused on how the network structure can affect people's opinions within a community and how high centrality nodes affect the overall COVID-19 vaccination acceptance rate. It highlighted the importance of the node position—the centrality measure in a network and the network structure when sentiment analysis is concerned—and demonstrated that social media influencers are not necessarily those with a high follower count.

This study presents a proof-of-concept experiment, revealing that applying social network analysis tools and techniques to Twitter data can help mine a wealth of information about the key actors, network dynamics, and communication patterns. Such knowledge can assist global health organizations, policymakers, government officials, and healthcare practitioners in having a realistic view of the extent of vaccine mistrust. Thus, our findings will support planning to address such challenges early before the crises intensify and affect global and local economies and humanity.

6. CONCLUSION AND FUTURE WORK

Community detection is a prevalent task in the field of social network analysis. It assists in visualizing the communication between participants and may identify key players in the interactions. This study analyzed and visualized data from tweets about the COVID-19 vaccine. Then, the VADER analyzer, a social media-based sentiment analysis tool, was used to classify the tweets as negative, neutral, or positive. Subsequently, a community detection algorithm was utilized to investigate which nodes affected the overall community opinion. Data were collected from Twitter using the Netlytic platform.

The study aimed to answer two research questions; the first question: How can we extract communities formed around COVID-19 vaccine discussions? The answer to this question was provided by constructing an interaction network and applying the Girvan-Newman community detection algorithm. The second research question: How do high centrality nodes in a community influence other users' opinions within the community? This question was answered by showing that the high centrality nodes' sentiment and the community's average sentiment are similar. The study findings emphasize the importance of the underlying structure of the social network for detecting communities and influencers. Adding the sentiment dimension enriches the results by showing the community's attitude regarding the COVID-19 vaccine and how high centrality nodes affect their communities.

The tweets analyzed in this study were gathered between November 11 and December 11, 2021, when there was active debate regarding vaccine distribution, dosages, and side effects. Nevertheless, we acknowledge that more accurate results could be obtained from an extended data collection period. Another limitation of the current work was using a single community detection algorithm—the Girvan–Newman algorithm; other available algorithms may yield better performance. Moreover, text cleaning and pre-processing could be expanded to enhance the sentiment results, such as dealing with emojis, spelling mistakes, and stemming.

More applied research based on the Twitter social media platform can be conducted to reveal hidden patterns of interactions, communities, and other results, including descriptive statistics. Additionally, such analyses would be more significant for authorities and governments when integrated with geographical locations and demographics that can be easily obtained from Twitter. Regarding constructing the social network, experimenting with adding edge weights and node attributes, such as the number of followers and age of the account, would be a good starting point for future research. A similar methodology (using Twitter data) could be used in other fields to understand the public's opinions about a specific topic. This can assist governments and organizations in decision-making and planning to mitigate people's resistance and apply new regulations or solutions. Additionally, longitudinal research can be conducted in the future to understand how people's online opinions change over time and how such changes are reflected in real life.

In sum, the study findings emphasize that real influencers can be found by carefully analyzing the interaction networks between users in a particular domain or topic, not by looking solely at the number of followers of Twitter users.

STATEMENTS AND DECLARATIONS

Acknowledgments: The authors would like to express their great appreciation to King Abdulaziz University for their generous support and help in offering the digital library resources needed to conduct this research. The authors would also like to thank the Faculty of Computing and Information Technology for supporting the faculty's academics, students, and research communities.

Funding: The authors did not receive support from any organization for the submitted work.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Competing interests: The authors have no competing interests to declare relevant to this article's content.

Research involving Human Participants and/or Animals: The current study did not involve animal or human subjects, so ethics approval is not applicable.

Informed Consent: The current study did not involve human subjects, so informed consent is not applicable.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1]. World Health Organization, Global Health Observatory. 2022.

[2]. Zhang J, Wang Y, Shi M, Wang X. Factors Driving the Popularity and Virality of COVID-19 Vaccine Discourse on Twitter: Text Mining and Data Visualization Study. JMIR Public Heal Surveill [Internet]. 2021 Dec 3;7(12):e32814. Available from: https://publichealth.jmir.org/2021/12/e32814

[3]. Easley D, Kleinberg J. Networks, Crowds, and Markets. Networks, Crowds, and Markets. 2010.

[4]. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci U A. 2002;99(12):7821–6.

[5]. Chaabani Y, Akaichi J. Bees colonies for terrorist communities evolution detection. Soc Netw Anal Min [Internet]. 2022;12(1):1–15. Available from: https://doi.org/10.1007/s13278-021-00835-y

[6]. Chaudhary L, Singh B. Community detection using unsupervised machine learning techniques on COVID-19 dataset. Soc Netw Anal Min [Internet]. 2021;11(1):1–9. Available from: https://doi.org/10.1007/s13278-021-00734-2

[7]. Mahmud S, Mohsin M, Hossain S, Islam MM, Muyeed A. The Acceptance of COVID-19 Vaccine: A Global Rapid Systematic Review and Meta-Analysis. SSRN Electron J [Internet]. 2021; Available from: https://www.ssrn.com/abstract=3855987

[8]. Lyu JC, Han E Le, Luli GK. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. J Med Internet Res [Internet]. 2021 Jun 29;23(6):e24435. Available from: https://www.jmir.org/2021/6/e24435

[9]. Eibensteiner F, Ritschl V, Nawaz FA, Fazel SS, Tsagkaris C, Kulnik ST, et al. People's Willingness to Vaccinate Against COVID-19 Despite Their Safety Concerns: Twitter Poll Analysis. J Med Internet Res [Internet]. 2021 Apr 29;23(4):e28973. Available from: https://www.jmir.org/2021/4/e28973

[10]. Xie Z, Wang X, Jiang Y, Chen Y, Huang S, Ma H, et al. Public Perception of COVID-19 Vaccines on Twitter in the United States. medRxiv Prepr Serv Heal Sci [Internet]. 2021 October 18; Available from: http://www.ncbi.nlm.nih.gov/pubmed/34704100

[11]. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing Public Opinion Towards COVID-19 Vaccines With Twitter Data in the United States: Spatiotemporal Perspective. J Med Internet Res [Internet]. 2021 Sep 10;23(9):e30854. Available from: https://www.jmir.org/2021/9/e30854

[12]. Griffith J, Marani H, Monkman H. COVID-19 Vaccine Hesitancy in Canada: Content Analysis of Tweets Using the Theoretical Domains Framework. J Med Internet Res [Internet]. 2021 Apr 13;23(4):e26874. Available from: https://www.jmir.org/2021/4/e26874

[13]. Bonnevie E, Goldbarg J, Gallegos-Jeffrey AK, Rosenberg SD, Wartella E, Smyser J. Content Themes and Influential Voices Within Vaccine Opposition on Twitter, 2019. Am J Public Health [Internet]. 2020 Oct;110(S3): S326–30. Available from: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2020.305901

[14]. Cossard A, Morales GDF, Kalimeri K, Mejova Y, Paolotti D, Starnini M. Falling into the Echo Chamber: the Italian Vaccination Debate on Twitter. 2020 March 26; Available from: http://arxiv.org/abs/2003.11906

[15]. Niu Q, Liu J, Kato M, Shinohara Y, Matsumura N, Aoyama T, et al. Public Opinion and Sentiment Before and at the Beginning of COVID-19 Vaccinations in Japan: Twitter Analysis (Preprint). JMIR Infodemiology. 2021;2:1–13.

[16]. Singh M, Jakhar AK, Pandey S. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc Netw Anal Min [Internet]. 2021 Dec 19;11(1):33. Available from: https://link.springer.com/10.1007/s13278-021-00737-z

[17]. Elbagir S, Yang J. Sentiment Analysis on Twitter with Python's Natural Language Toolkit and VADER Sentiment Analyzer. In: IAENG Transactions on Engineering Sciences [Internet]. WORLD SCIENTIFIC; 2020. p. 63–80. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789811215094_0005

[18]. Loria S. Textblob Documentation. Release 0.16. 2018.

[19]. Ferdous Z, Akhter R, Tahsin A, Nuha Mustafina S, Tabassum N. Sentiment Analysis on COVID-19 Vaccine Twitter Data using Neural Network Models. In: Proceedings of the 2nd International Conference on Computing Advancements [Internet]. New

York, NY, USA: ACM; 2022. p. 435–41. Available from: https://dl.acm.org/doi/10.1145/3542954.3543064

[20]. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 Vaccine Hesitancy: Text Mining, Sentiment Analysis and Machine Learning on COVID-19 Vaccination Twitter Dataset. Expert Syst Appl [Internet]. 2022 Sep;118715. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417422017407

[21]. Bhatnagar S, Choubey N. Making sense of tweets using sentiment analysis on closely related topics. Soc Netw Anal Min [Internet]. 2021;11(1):1–11. Available from: https://doi.org/10.1007/s13278-021-00752-0

[22]. Parés F, Garcia-Gasulla D, Vilalta A, Moreno J, Ayguadé E, Labarta J, et al. Fluid Communities: A Competitive, Scalable and Diverse Community Detection Algorithm. 2017 March 27; Available from: http://arxiv.org/abs/1703.09307

[23]. Kwok SWH, Vadde SK, Wang G. Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. J Med Internet Res [Internet]. 2021 May 19;23(5):e26953. Available from: https://www.jmir.org/2021/5/e26953

[24]. Gruzd A. Netlytic: Software for Automated Text and Social Network Analysis [Internet]. Available from: http://netlytic.org

[25]. 25. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA: O'Reilly Media; 2019.

[26]. Alhajj R, Rokne J. Sociograph Representations, Concepts, Data, and Analysis. In: Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY; 2018. p. 2839–41.

[27]. Bastian M, Heymann S, Jacomy M. Gephi : An Open Source Software for Exploring and Manipulating Networks. In: Third International AAAI Conference on Weblogs and Social Media. 2009. p. 361–2.