Social Network Analysis of Twitter User Conversations About Educational Technology Companies Using Network Properties and Centrality Metrics

Bart Simpson and Homer Simpson
Department of Nuclear Power Engineering
University of Springfield
Springfield, Nostate 12345, USA
{bart.simpson & homer.simpson}@uspringfield.edu

Monkey King, Bajie Zhu and Seng Tang
Department of Intelligent Robotics
University of Huaguoshan
Huaguoshan, Jileshijie Province, China
monkey.king@uhuaguoshan.edu.cn

***Abstract* - The number of social media users in Indonesia increases by 10 million users or 6,3 percent between 2020 and 2021. This phenomenon could open new opportunities for companies to increase the effectiveness of their marketing activities on social media, including for companies in the educational technology (edtech) sector. Ruangguru and Zenius are the most dominant edtech companies in the edtech industry competition in Indonesia, both from the perspective of investors, and from the number of followers on Twitter. Twitter can be used by companies as a marketing medium to reach a wider audience than the competing companies. Therefore, companies need to know how the company's marketing activities or brand recognition on Twitter compared to the competing companies. This study uses the application of Social Network Analysis (SNA) in analyzing the social network of Twitter users' conversations about the two edtech companies. The SNA metrics used in this study are network properties and centrality metrics. Based on the results obtained in the calculation of network properties and centrality metrics, the social network structure of Twitter users' conversations about Zenius is superior to Ruangguru, with a user who has the most influence on the flow of information dissemination on each network is dominated by Schfess account. This shows that on the Zenius network, the distribution of information is more efficient than on the Ruangguru network.**

Index Terms* – *Centrality, Educational Technology, Network Properties, Social Network Analysis.

I. Introduction

The increasing use of the internet in Indonesia opens up opportunities for the emergence of innovation in the use of information, communication, and technology in the education sector, namely in the form of educational technology (edtech) platforms. The edtech sector in Indonesia is still in its growth phase, with almost all companies experimenting with their products or markets. Based on data from the World Bank, the establishment of edtech companies has also coincided with the increase in internet penetration in Indonesia [1].

Two educational technology (edtech) companies, namely Ruangguru and Zenius, are included in Indonesia's list of edtech companies that have been dominant in terms of user growth and investor attention over the past few years [1]. Ruangguru and Zenius are also edtech platform accounts in Indonesia that have the most followers on the online social networking site, Twitter. However, having many Twitter followers does not mean much if the level of interaction is low and the company does not involve these followers in marketing activities. Therefore, an analysis is needed that can help companies understand their consumer interaction patterns on Twitter so that companies can identify how their product marketing activities compare to competing companies.

One of the methods in social media analytics that is commonly used to analyze the pattern of interaction between individuals is Social Network Analysis (SNA). SNA is an analytical approach that utilizes graph theory to identify the structure of a social network. Some social networks may consist of Twitter users, denoted by a node, and the interaction between Twitter users, represented by an edge.

There are several previous studies that discuss the implementation of SNA in the formulation of marketing strategies on social media. First, research was conducted by Ioannis Antoniadis and Anna Charmantzi on the application of SNA in building communication and branding strategies by building social capital on social networking sites [2]. Second, the research is written by Arnaldo Litterio et al. regarding the application of SNA in marketing to identify opinion leaders [3]. Third, research is written by Itai Himelboim and Guy Golan on the approach of social networks in analyzing the role of influencers in viral advertising [4].

Based on the background described above, the author conducted a study on social network analysis of Twitter user conversations about two edtech companies, namely Ruangguru and Zenius. The metrics used in the SNA approach are network properties and centrality metrics.

II. Research Methodology

A. *Research Data*

The data used in this study is tweet data obtained from Twitter with the keywords "Ruangguru" and "Zenius" from July 1 to September 30, 2021. The data collected is 39,219 tweets with 5,488 users and 4,982 conversations for Ruangguru and 2,605 users and 2,123 conversations for Zenius.

B. *Data Collection*

Data collection is done by scraping from Twitter. This step is carried out using Python and the Twint library. The data that is scraped is data that can be freely accessible by the public.

C. *Data Pre-processing*

All tweets collected are then carried out in the data cleaning stage to remove irrelevant data and make the analysis easier.

Next, the data transformation stage is carried out in the form of an edge list with the help of the Pandas and Networkx packages from the Python programming language.

D. *Social Network Analysis*

Social Network Analysis (SNA) is an analytical approach that can identify the structure of social networks by utilizing graph theory. The formed social network can consist of users, denoted by nodes, and interactions between users, represented by sides. SNA is necessary because it brings new opportunities to understand individuals or groups regarding their interaction patterns. SNA uses a graph theory approach to describe the network model's structure. In this study, the authors divide the measurement into two parts, namely network properties and centrality [5].

1. *Network Properties*

Each network model that has been processed through the pre-processing data stage has several properties whose values will be calculated as follows:

1) *Order and Size*

A network order is the number of nodes, and size is the number of edges on the network. The number of orders and sizes in a social network shows the number of users interacting. In the following discussion, the number of vertices will be represented as a variable *n*, and the number of edges will be defined as a variable *m* [5].

$$m = |E(G)|,\ n = |V(G)| \tag{1}$$

2) *Density*

Density is a calculation of the number of sides compared to the maximum number of sides in a network. Density describes the density of the network. The higher the density value, the better because it illustrates that users on the network are more connected to each other. The formula for calculating density is as follows:

$$\rho(G) = \frac{m(G)}{m_{max}(G)} = \frac{m}{n(n-1)/2} \tag{2}$$

with *m* represents the number of edges in the network, and *n* represents the number of nodes in the network [5].

3) *Modularity*

Modularity is a metric used to determine the quality of network division into groups. Maximizing the value of modularity efficiently can be done with the Louvain algorithm. Louvain's algorithm is a community detection algorithm that recursively combines groups into one node and executes modularity clustering on the summarized network. Louvain's algorithm consists of two stages: Modularity Optimization and Community Aggregation. Changes in the value of modularity (ΔQ) can be calculated by the following formula [5]:

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{3}$$

descriptions:

$\sum_{in}$: number of sides in Group *C.*

$\sum_{tot}$: the number of edges attached to the vertices in Group *C.*

k_i : number of edges connected to vertex *i.*

$k_{i,in}$: the number of edges in vertex *i* that are connected to vertices in Group *C.*

m : the number of edges in the network.

The number of groups will continue to decrease with each iteration or pass. The pass is repeated until there are no more changes and maximum modularity is reached [5].

4) *Diameter*

Diameter is the longest shortest-path distance between a pair of nodes in the network. The smaller the diameter value, the better because the disseminating information between a user and another user, with the furthest distance, only needs to pass through a few users.

$$D = max\left\{d_{v_i,v_j} : v \in V\right\} \tag{4}$$

with *V* represents the set of nodes in the network, and d_{v_i,v_j} represents the shortest-path distance between vertices *i* and *j*. In large networks, the shortest-path can be determined using the Breadth-First Search (BFS) algorithm [5].

5) *Average Path Length*

The average path length calculates the average shortest-path distance between each pair of nodes in a network. The smaller the value of the average path length, the better because the average distance that must be traveled to disseminate information is shorter. The formula for calculating the average path length is as follows.

$$l = \frac{2}{n(n-1)} \sum_{i \neq j} d_{v_i,v_j} \tag{5}$$

with *n* represents the number of nodes in the network, and d_{v_i,v_j} represents the shortest-path distance between vertices *i* and *j* [5].

6) *Average Degree*

The average degree calculates the number of edges connecting a node to another node on a network. The greater the average degree value possessed by the network, the better because if a user disseminates information to other users, it will accelerate the dissemination of information in the network. The formula for calculating the average degree is as follows:

$$k = \frac{1}{n} \sum_{i=1}^{n} k_i \tag{6}$$

with *n* represents the number of nodes in the network, and k_i represents degrees at node *i* [5].

7) *Connected Components*

Connected components are parts that are separate or not connected to the entire network. The smaller the value of connected components, the better because users are not separated too much in small groups that are not connected. To identify the number of connected components can use the BFS algorithm [5].

2. *Centrality*

After the network properties metric, the following metric is centrality. Centrality measurement aims to identify a network's most influential users (key actors). There are four measurements of centrality in this study, as follows:

1) *Degree Centrality*

The degree centrality metric describes the size of a user's social connections on the network. A node with a high degree of centrality may have a central position in the network but may also be far away from the edge of the network. Here is the formula for the degree centrality for node *i*:

$$C_D(i) = k_i = \sum_{i\neq j}^{n} a_{v_i v_j} \quad (7)$$

$$a_{v_i v_j} = \begin{cases} 1, & \text{if there is an edge between vertices } i \text{ and } j \\ 0, & \text{other} \end{cases}$$

with *n* represents the number of nodes in the network [6].

2) *Betweenness Centrality*

Betweenness centrality is a metric not concerned with how many social connections a user has but where the user is placed in the network. To calculate the betweenness centrality value at node *i*, we can calculate the proportion of the shortest path between node *j* and *h* that passes through node *i* [6].

$$C_B(i) = g(i) = \frac{2}{(n-1)(n-2)} \cdot \sum_{h\neq i, h\neq j, j\neq i}^{n} \frac{\sigma_{hj}(i)}{\sigma_{hj}} \quad (8)$$

descriptions:

$\sigma_{hj}(i)$: the number of shortest-paths between node *h* and node *j* that pass-through node *i*.

σ_{hj} : the number of shortest-paths between node *h* and node *j*.

n : the number of nodes in the network.

3) *Closeness Centrality*

Closeness centrality is a calculation to find the node closest to all other nodes in a network. Closeness centrality for a node is the inverse of the average shortest-path distance from that node to every other node in the network. The formula for calculating closeness centrality at node *i* is as follows:

$$C_c(i) = \frac{n-1}{\sum_{i\neq j} d_{v_i, v_j}} \quad (9)$$

with *n* represents the number of nodes in the network, and d_{v_i, v_j} represents the shortest-path distance between vertices *i* and *j* [6].

4) *Eigenvector Centrality*

Eigenvector centrality is a measure that considers the number of connections of a user in the network. In other words, this metric considers the centrality of the node itself as well as the nodes connected to it. Intuitively, this measure takes into account not only how many users are known but also who is known.

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{i,j} \cdot x_j \quad (10)$$

$$a_{i,j} = \begin{cases} 1, & \text{if there is an edge between vertices } i \text{ and } j \\ 0, & \text{other} \end{cases}$$

descriptions:

$a_{i,j}$: adjacency matrix

x_j : centrality value of node *j*

λ : largest eigenvalue

The above formula can be written in vector notation as an eigenvector equation as follows:

$$(A - \lambda I)x = 0 \quad (11)$$

To calculate this metric, we need the largest eigenvalue and eigenvector of the adjacency matrix. To find eigenvalues and eigenvectors, we can use the characteristic equation of polynomials [6].

E. *Network Model Visualization*

The edge list data is then processed using the Gephi application to visualize the network model based on the metrics calculated in the previous stage.

III. Analysis & Discussion

A. *Data Collection*

Data retrieval is done by the scraping method. The tweet data was scraped from Twitter with the search keywords "ruangguru" and "zenius" from July 1, 2021, to September 30, 2021. This stage produces an output in the form of a CSV file containing scraped tweet data. The size of the raw data used in this study was 39,219 rows and six columns.

B. *Data Pre-processing*

The stages of pre-processing data in this study are as follows:

1. *Deleting Duplicate Tweets*

Tweet data that has been collected at the scraping stage allows there to be duplicate data. Therefore, one of the duplicate data must be removed until each data to be analyzed is unique data.

2. *Deleting Tweets That Have No Interaction*

This stage is carried out because tweets that do not have interaction can be considered self-loop nodes. Thus, deleting tweets with these criteria will facilitate the subsequent analysis phase.

3. *Retrieving Conversation Tweets Between Users*

The author will delete tweets directly related to the official company accounts of Ruangguru and Zenius. Thus, it will only retrieve tweets that are conversations between regular users. This stage is carried out because the tweets that will be analyzed are conversational interactions only between users.

4. *Grouping Tweets About Edtech Companies*

Before analyzing each social network formed, the data frames were first grouped based on the context of the conversation about Ruangguru or Zenius. This stage is carried out because social network analysis will be conducted on each social network formed regarding Ruangguru and Zenius.

5. *Transform Data to Edge List Form*

The final step at this stage is to transform the two data frames into edge lists. An edge list is a simple representation of a graph. Forming an edge list takes at least two nodes representing the name of the account that replies to a tweet and the name of the account that is responded to. The size of the two edge list data generated at this stage is 5,231 rows, three columns for Ruangguru and 2,156 rows, and three columns for Zenius.

C. *Social Network Analysis*

The next step is to process edge list data for Ruangguru and Zenius using the Social Network Analysis (SNA) approach. The graph without direction was chosen because, in this study, the author only focuses on analyzing the distribution of information based on the presence or absence of interaction between a node and another node so that the direction of exchange or the order of nodes on the edge list is not included in the focus of this study.

1. *Calculation of Network Properties Metrics*

TABLE I
NETWORK PROPERTIES METRICS

Network Properties	Ruangguru	Zenius
Size	4,982	2,123
Order	5,488	2,605
Density	.00033089	.00062594
Modularity	.87334	.88822
Diameter	19	13
Average Path Length	5.2505	3.9469
Average Degree	1.8156	1.6299
Connected Components	1,022	587

Table 1 compares all calculation results of network properties metrics on the Ruangguru and Zenius networks. It can be seen that the structure of the Twitter user conversation network regarding Ruangguru only excels in three categories, namely in size, order, and average degree metrics. Meanwhile, the network structure of Twitter users' conversations about Zenius excels in five categories: density, modularity, diameter, average path length, and connected components metrics.

2. *Calculation of Centrality Metrics*

TABLE 2
CENTRALITY METRICS ON THE RUANGGURU NETWORK

User	DC	BC	CC	EC
	Score/Rank	*Score/Rank*	*Score/Rank*	*Score/Rank*
schfess	.922 / 1	.132 / 1	.182 / 1	.612 / 1
subschfess	.638 / 2	.0897 / 2	.170 / 3	.286 / 2
ambisfs	.603 / 3	.0560 / 6	.146 / 51	.137 / 3
sbmptnfess	.0454 / 4	.0588 / 5	.165 / 5	.121 / 4
guidance204	.0191 / 5	.0262 / 9	.164 / 6	.0725 / 5

Table 2 shows all the centrality metric values in the Ruangguru network. From the four metrics, it can be seen that users with account names Schfess and Subschfess are users who always occupy the top three.

TABLE 3
CENTRALITY METRICS ON THE ZENIUS NETWORK

User	DC	BC	CC	EC
	Score/Rank	*Score/Rank*	*Score/Rank*	*Score/Rank*
schfess	.141 / 1	.143 / 1	.206 / 1	.694 / 1
sbmptnfess	.0791 / 2	.0767 / 2	.177 / 8	.0661 /3
subschfess	.0710 / 3	.0733 / 3	.173 / 9	.109 / 2
sabdaps	.0188 / 4	.0179 / 7	.132 / 496	.0024 / 829
zenius_oliv	.0184 / 5	.0396 / 4	.197 / 2	.0543 / 4

Table 3 shows all the centrality metric values on the Zenius network. From the four metrics, it can be seen that users with the account name Schfess are users who always occupy the top position. This shows that the account is a key actor in disseminating information on each network.

D. *Network Model Visualization*

At this stage, the author will display a visual representation of the network model that has been processed in the previous step to make it easier for readers to identify network analysis results. Two attributes on the network will assist in processing the visualization, namely the characteristics of the modularity and betweenness centrality metrics. The modularity metric is needed at the visualization stage because it can show the groups or clusters formed on the network. The betweenness centrality metric is required at the visualization stage because it can show the key actors. They are central in disseminating information in each group and within the network.

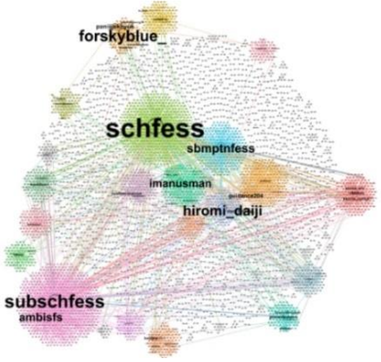

Fig. 1 Ruangguru Network Visualization

Figure 1 is a visualization of the Ruangguru network. There are 20 large groups in the network (groups that are colored). A large group is a group that has a percentage of members more than or equal to 1 percent.

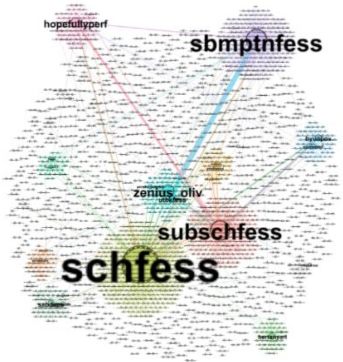

Fig. 2 Zenius Network Visualization

Figure 2 shows the network model of Twitter users' conversations about Zenius. There are 11 large groups or clusters in the network.

IV. Conclusions & Recommendations

Based on the analysis and research results in the previous chapter, it can be concluded that:

1. Based on the analysis and comparison of network properties metrics, the social network structure characteristics of Twitter users' conversations about Zenius excel in five of the eight network properties metrics: density, modularity, diameter, average path length, and connected components. Meanwhile, the characteristics of the social network structure of Twitter users' conversations about Ruangguru only excel in three of the eight network property metrics: size, order, and average degree. In the Ruangguru network, the social network structure characteristics are superior in measuring connection type but not exceptional in measuring distribution and segmentation types. This shows that although the Ruangguru network has many actors and conversational interactions, the distribution of information and the intensity of conversations between these actors is not better or more efficient than on the Zenius network.
2. Based on the analysis and comparison of all centrality metrics (degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality) owned by actors in the Ruangguru and Zenius network. It can be identified that the key actor in each of these networks is the Schfess account, which is an account of the student community throughout Indonesia.

Based on the results of the analysis and research in the previous chapter, there are several suggestions for companies:

1. The company's Twitter account can be more active in engaging with followers so that the conversations of Twitter users about the company not only have an increasing trend but can also form a social network with a superior structure compared to competing companies.
2. Companies can also collaborate with key actors able to spread information or brand recognition more quickly and widely on Twitter. Thus, the company can reach a larger audience than competing companies.
3. Companies can also include stakeholders in interacting with Twitter users. Users with a reasonably high centrality metric value on each network are users with the account names Sabdaps and Imanusman. After further investigation, the two accounts are the founders of Zenius and Ruangguru. This shows that many Twitter users interact with people who represent the company.

References

[1] Bhardwaj, R.; Yarrow, N.; dan Cali, M. 2020. "EdTech in Indonesia: Ready for Take-off". Washington DC: World Bank.

[2] Antoniadis, I. dan Charmantzi, A. 2016. "Social Network Analysis and Social Capital in Marketing: Theory and Practical Implementation". *International Journal of Technology Marketing* 11(3): 344-359.

[3] Litterio, A. M.; Nantes, E. A.; dkk. 2017. "Marketing and Social Networks: A Criterion for Detecting Opinion Leaders". *European Journal of Management and Business Economics* 26: 347–366.

[4] Himelboim, I. dan Golan, G. J. 2019. "A Social Networks Approach to Viral Advertising: The Role of Primary, Contextual, and Low Influencers". *Social Media+ Society* 5 (3).

[5] Barabási, A. L. 2015. "Network Science". http://networksciencebook.com/.

[6] Fornito, A.; Zalesky, A.; Bullmore, E. 2016. "Fundamentals of Brain Network Analysis". USA: Academic Press Elsevier.