Register 6 (1) 2020 pp. 22-31
Jurnal Ilmiah Teknologi Sistem Informasi (Scientific Journal of Information System Technology)

ISSN 2502-3357 (Online)
ISSN 2503-0477 (Print)

Available online to www.journal.unipdu.ac.id

Unipdu

S2-Accredited – SK No. 34/E/KPT/2018
Journal page is available to www.journal.unipdu.ac.id:8080/index.php/register

Community detection in twitter based on tweets similarities in indonesian using cosine similarity and louvain algorithms

Akhmad Irsyad [a], Nur Aini Rakhmawati [b]

[a,b] Information System Department, Institut Inteknologi Sepuluh Nopember, Surabaya, Indonesia

email: [a] sentinel_irsyad@hotmail.com, [b] nur.aini@is.its.ac.id

ARTICLE INFO

Article history:
Received 19 June 2019
Revised 11 September 2019
Accepted 7 November 2019
Published 13 December 2019

Keywords:
community detection
Louvain algorithm
social network
text similarity
Twitter

IEEE style in citing this article:
A. Irsyad and N. A. Rakhmawati, "Community Detection in Twitter Based on Tweets Similarities in Indonesian using Cosine Similarity and Louvain Algorithms," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 6, no. 1, pp. 22-31, 2020.

ABSTRACT

Twitter is now considered as one of the fastest and most popular communication media and is often used to track current events or news. Many tweets tend to contain semantically identical information. When following an activity or news, sometimes in tweeting people do it in groups. Therefore, it is necessary to have a useful technique for grouping users based on the tweets similarities. In this study, cosine similarity method is used to examine the similarity of tweets between accounts, and a graph-based approach is proposed to detect communities. Graphs are first depicted from similarities between tweets and next community detection techniques are applied in graphs to group accounts that have similar tweets. The reason for using these two methods is that compared to other methods, the accuracy of cosine similarity is higher while Louvain can result a better modularity. From this research, it was concluded that cosine similarity and Louvain algorithm could be used in community detection on social media.

1. Introduction

Nowadays, the number of social media users is increasing rapidly. It is estimated that the number of Twitter users registered in 2016 has reached 317 million [1]. This shows that social media, especially Twitter has become an important communication media. Social media technology allows messages to be sent quickly and widely. It can also create a viral when the topic attracts public attention. Twitter has quickly become one of the most popular social network sites. It is not only used as a communication media, but is also used for exchanging information, advertising and campaigning for political parties [2].

The social networks development has become an interesting research object for researchers, one of which is community detection on very large and complex networks such as social networks [3]. Community detection aims to divide the network and described in the form of graph. If the entity has a correlation, it can be said to be one community. This community utilization can then be used for various purposes such as finding the market targets, ranking the popularity of a product, detecting issues in society, detecting terrorist networks, and many more [4]. The similarity of tweets will be used as a baseline for the communities formation in community detection.

The purpose of this research is to find community accounts or groups that discuss an event in

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

social media using cosine similarity and Louvain algorithm displayed in the form of graphs to make it easier to understand the community formation and find pattern of the community. Cosine Similarity will be used to determine the similarity between tweets while Louvain algorithm is utilized to detect communities in networks. Based on previous studies it was found that compared to other algorithm, cosine similarity is advantageous in term of high degree of accuracy [5] while Louvain's algorithm can detect communities with high modularity better and process faster [6].

This paper is organized as follow. In the first section we provide background information about the community detection problem, second section briefly introduces previous work on text similarity and community detection, the proposed methodology is explained in Section three, the results of the study are explained in Section four. And the last section concludes the work of this paper.

2. Research Method

There are some studies relating to the similarity of tweets and community detection. Lazuardi research [7] collected data from Twitter and processed it using Text Mining and Social Network Analysis. This research applied Association Rules calculations to find frequently used words and collections of words about perceptions of a company's brand quality. Louvain algorithm was used to optimized the modularity to find word groups. The results of this study is discovering the most dominant structure of words regarding to the perception of brand quality.

Nur et al. [2] used data from Twitter which was then processed using Text Mining and Social Network Analysis. The first step was calculating the similarity between tweets using cosine similarity. It measured the level of similarity between users based on their interactions. The method used to determine whether users were in the same community was Genetic Algorithm using some twitter features; following, follower, mention and reply as variables.

Dutta et al. [8], to measure the similarity between tweets, researchers consider not only the existence of general terms (hashtags and URLs), but also the semantic similarities between tweets. The WordNet tool is used to capture semantic similarities between tweets that might use different terms to express the same information. Using community detection techniques, the same tweets are clustered and representative tweets are chosen from each cluster (from the same tweet) to be included in the summary.

Fócil-Arias et al. [9] conducted a study to analyze Content, Microblog search, and TimeLine illustrations using the cosine method which compare the similarities between two space vectors per tweet. The vector was obtained from the Bag-of-Words and word2vec approaches by calculating the similarity of tweets using cosine similarity with Microblog Cultural Contextualization 2017 workshop dataset. The result was to determine the relevance of tweets according to each event from four European festivals: Charrues, Transmusicales, Avignon and Edinburgh.

Conover et al. [10] examined the political communication network on the Twitter service for six weeks before the election in the United States in 2010. Taking data from the Twitter 'gardenhose' API, he successfully identified 250,000 politically relevant messages (tweets) sent by more than 45,000 users. Users is called connected in twitter when one re-broadcast twitter content sent by another. It is detected using the jaccard method. They are also said to be connected when one mention or re-tweet others in a post. The research shows that the retweet network points a very modular structure, clustering users into two homogeneous communities, the left and right politics. Instead, it is found that the networks did not exhibit this kind of political segregation, which resulted in users being exposed to individuals and information they would not have chosen before [10].

3. Research Methodology

This section demonstrates the design of the system flow that will be implemented in this research which can be seen in Figure 1. The first step is collecting dataset from twitter.

3.1. Collecting data

Twitter is a social media that connects everyone in the world and allows its users to communicate with each other in a short 140-character message called a tweet. Its users can keep up with the latest news about the topics they are interested in. It provides access to Twitter data through an API that enable developers to build and expand their applications based on their own creativity. It continues to develop

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

the API so that it experience growth [11] . This study uses Twitter as the source of the dataset and collects tweets using the Twitter Streaming API. Tweets searching is conducted on March 23, 2019 based on FPI query.

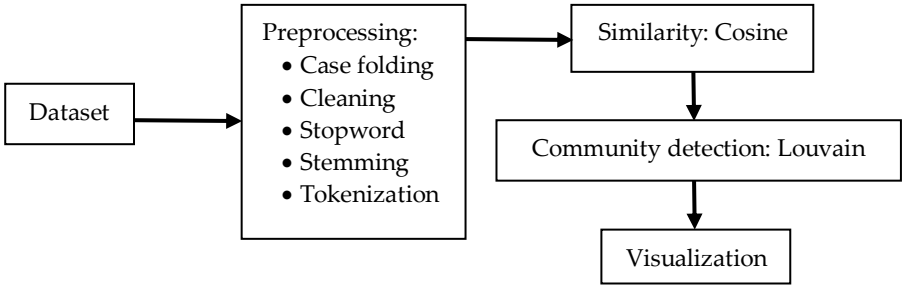

Figure 1. Proposed architecture

3.2. Preprocessing

Preprocessing is implemented to avoid incomplete data, data interruptions, and inconsistent data. Preprocess text stages in this study include [12]:

1. Removing URL (http://www.situs.com) and email (name@situs.com).
2. Removing Special Characters in Twitter, This process is performed by removing special characters such as hashtags, user names (@username), and special characters (for example RT, which indicates that the user is retweeting something).
3. Removing Symbols. This step is done to remove the symbols and punctuations in the tweet.
4. Removing Stopwords. Stopwords are words that do not affect the classification process.
5. Stemming. The purpose of stemming is to get the basic (word stem) of a word. Since words derived from semantics tend to be similar to their basic words (word stem), the process of stemming algorithm is often used in accordance with the language to be studied [13]. Stemming is performed using Nazief and Adriani Algorithm [14].
6. The tokenization process is the process of cutting the input string based on its each letter. An example of tokenization method is N-gram. N-gram is a probability model that was originally designed by Russian mathematicians in the early 20th century and then developed to predict the next word or character in a string sequence [15]. Strings can be in the form of characters or words according to the needs of the application. In this study char unigram is used as the tokenization method.

3.3. Cosine similarity

The Cosine Similarity method is one method to measure the level of similarity between two strings. At this stage the similarity measurement of the tweet content from preprocessing stage is conducted. The advantage of the this method is that it is not affected by the sequence and length of the string, while the weakness is that it cannot distinguish the meaning of a word [16]. In this process the similarity of each tweets are checked using cosine similarity equation (Equation (1)),

$$similarity(A,B) = \frac{A \times B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \times \sum_{i=1}^{n} B_i^2}} \quad (1)$$

$\boldsymbol{A_i}$ = Number of occurrences of the i-index word from the list of words in sentence A.

$\boldsymbol{B_i}$ =Number of occurrences of the i-index word from the list of words in sentence B

3.4. Community detection in graphs

Community Detection in graphs aims to find communities based on network structure, for example, to find densely connected groups of nodes. Similar types of nodes in a network form a community. The edge that connects the nodes in the community is Intra-community edge, while that connect the nodes in different communities are called edges Inter-community [8]. Figure 2 points the Intra-community and the Inter-community edges between different communities.

In this study, we made a graph where the nodes represent twitter usernames and edges reflect the similarities between tweets. We then detect the community in the graph based on the similarity of

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

the tweets based on Louvain algorithm. To perform this algorithm we use the neo4j application [17]. The calculation equation of the community detection Louvain Method is [18] using Equation (2),

$$Q = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - \frac{K_i k_j}{2m}\right]\delta(c_i c_j) \quad (2)$$

In Community Detection in graphs we use closeness centrality to measure how central a node in its cluster. Closeness Centrality is a way of detecting nodes that can spread information efficiently through graphics. The node with the shortest distance to other nodes is assumed to be the central node of the cluster [17].

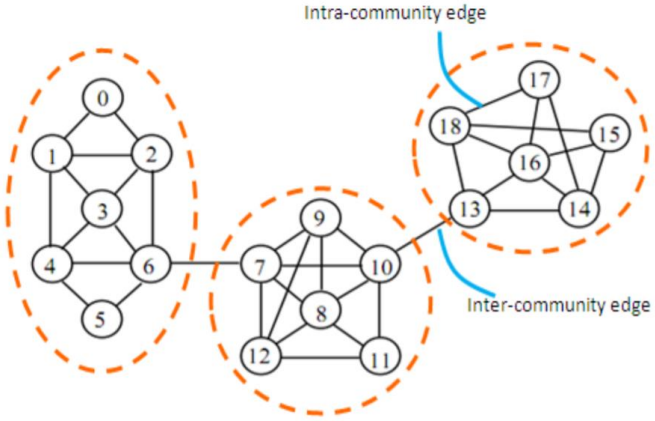

Figure 2. Community Structure in graph shows intra-community edges and inter-community e**dges [8]**

3.5. Closeness centrality

Closeness centrality is a basic indicator of complex networks. Closeness centrality has an important role in the study of complex networks, especially in identifying the importance of nodes in a network. Closeness centrality depends on the length of the path from one node to all other nodes in the network. It is defined as the inverse of the total length [19]. Equation (3) is the Closeness centrality equation,

$$\mathrm{Ci} = \frac{N-1}{\sum_j d(i,j)} \quad (3)$$

4. Result and Discussion

4.1. Collecting data

Tweets were collected on May 9, 2019, coincided with the Indonesian presidential election process. Based on FPI search queries, 10000 accounts that have tweeted related to FPI were obtained. From those 10000 accounts, crawling data was performed with 10 maximum of the last tweets from each account resulting 77196 tweets from 7817 Twitter accounts which were then preprocessed in the next step.

4.2. Preprocessing

In the pre-processing stage, 9384 tweets from 991 Twitter accounts were obtained in Table 1.

Table 1. Preprocessing result example

Before Preprocessing	After Preprocessing
@Lutfi_fams: Menteri Dalam Negeri dan Rakyat Indonesia: Dukung FPI terus eksis - Sign the Petition! https://t.co/qiDJxYyoQv lewat @ChangeOrg_ID	@Lutfi_fams: menteri dalam negeri rakyat indonesia dukung fpi terus eksis sign the petition lewat

4.3. Cosine similarity

In this stage, 9384 tweets were compared one by one. From the comparation, we filter 66746 pairs of tweets that has similarities above 0.9. Cosine similarity is chosen because compared to other method, it has a large similarity value, that allows more data can be obtained fro the sam source. For instance, based on the example in Table 2, the data do not considered to be the same although it distinguishes only one letter.

Each account will be considered to be in the same community if each account has the same tweet content of at least 5 tweets. As can be seen in Table 3 both accounts have the same five tweets, so both

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

will be considered to be in the same community. After conducting cosine similarity, the similarity on each account will be measured. When two accounts have at least five pairs of tweets that have similarity above 0.9, then these two accounts are considered to be in the same community. The data that has been processed then entered into neo4j to make community detection using the louvain algorithm which resulting five main communities as contained in Table 4.

Table 2. Cosine similarity result

Tweet 1	Tweet 2	Similarity	Methods
Aishara05108168: kubu02 emng pcundank sejati diskualifikasi dri kontestasi	Agraroberto8: kubu02 emg pcundank sejati diskualifikasi dr kontestasi	0.982299486257503	*Cosine similarity*
		0.7746478873239436	*jaccard*

Table 3. Account example that form community

Account 1	Tweet 1	Akun 2	Tweet 2
GoSumutdotcom	Kalau Tidak Menang, Jangan Kalah Itu Pesan Dejan Untuk Madura United https://t.co/GQOg6GKZBp. #GoSumut"	GoRiauCom	Kalau Tidak Menang, Jangan Kalah Itu Pesan Dejan Untuk Madura United https://t.co/eI86i3a2bb. #GoRiau"
GoSumutdotcom	Satpolair Polres Tanjungbalai Â Bagi- Bagi Takjil di Perairan Asahan https://t.co/IQfHy2sCtX. . #GoSumut	GoRiauCom	Alfredo Bersyukur Bhayangkara FC Dapat Satu Poin di Tenggarong https://t.co/ySZ37XVs69. #GoRiau
GoSumutdotcom	Alfredo Bersyukur Bhayangkara FC Dapat Satu Poin di Tenggarong https://t.co/Bl7cU7qzOp. . #GoSumut	GoRiauCom	Aji Santoso Terapkan Filosofi Permainan sepakbola Indah https://t.co/98BwXWkdCA. #GoRiau
GoSumutdotcom	Aji Santoso Terapkan Filosofi Permainan Sepak Bola Indah https://t.co/ljUQQSpT3U. . #GoSumut	GoRiauCom	Jafri Sastra Soroti Finishing Dan Kepercayaan Diri Pemain PSIS https://t.co/a19VmXGyz5. #GoRiau
GoSumutdotcom	Jafri Sastra Soroti Finishing Dan Kepercayaan Diri Pemain PSIS https://t.co/zZb5S0Llug. . #GoSumut	GoRiauCom	Stefano Puas Bali United Raih Tiga Poin https://t.co/kZndyq0lkz. #GoRiau
GoSumutdotcom	Stefano Puas Bali United Raih Tiga Poin https://t.co/qNpBjtxbqt. . #GoSumut	GoRiauCom	Kesabaran Kunci Sukses Kalteng Putra Bungkam PSIS https://t.co/xnccH1iOY9. #GoRiau
GoSumutdotcom	"Pasca Aksi Mogok, Pelayanan di RSUD Sibuhuan Kembali NormalÂ https://t.co/GgV5SG1xbo . #GoSumut"	GoRiauCom	"Jelang Kualifikasi Piala Dunia 2020, Tinnas Jalani TC di Yogyakarta https://t.co/H8QsbFKpXg .. #GoRiau"
GoSumutdotcom	Kesabaran Kunci Sukses Kalteng Putra Bungkam PSIS https://t.co/akmPCyt9ts. #GoSumut	GoRiauCom	Timnas Indonesia U 16 Latihan Internal Game https://t.co/B0eKrGwmQQ #GoRiau
GoSumutdotcom	Jelang Kualifikasi Piala Dunia 2020, Tinnas Jalani TC di Yogyakarta https://t.co/xaukLQRnN. #GoSumut"		
GoSumutdotcom	Timnas Indonesia U 16 Latihan Internal Game https://t.co/13yTNEHCdw .#GoSumut		

Table 4. Main *community*

Community	Number
2	14
3	14
5	8
15	7
1	5

The visualization graph of the data can be seen in Figure 5, which represents 5 main colors resulted from community detection, consisting yellow represents community 1, dark purple represents community 15, light purple represents community 5, green represents community 3, and red represents community 2. Community detection uses Twitter account to find groups of users in the same community who is connected by edge.

Figure 3 shows several large user groups in the network. The Twitter account that is the center of the user group is represented by a large node, which also indicates that it has a large Closeness Centrality and that its user has a huge centrality on Twitter social media.

ISSN 2502-3357 (Online)
ISSN 2503-0477 (Print)

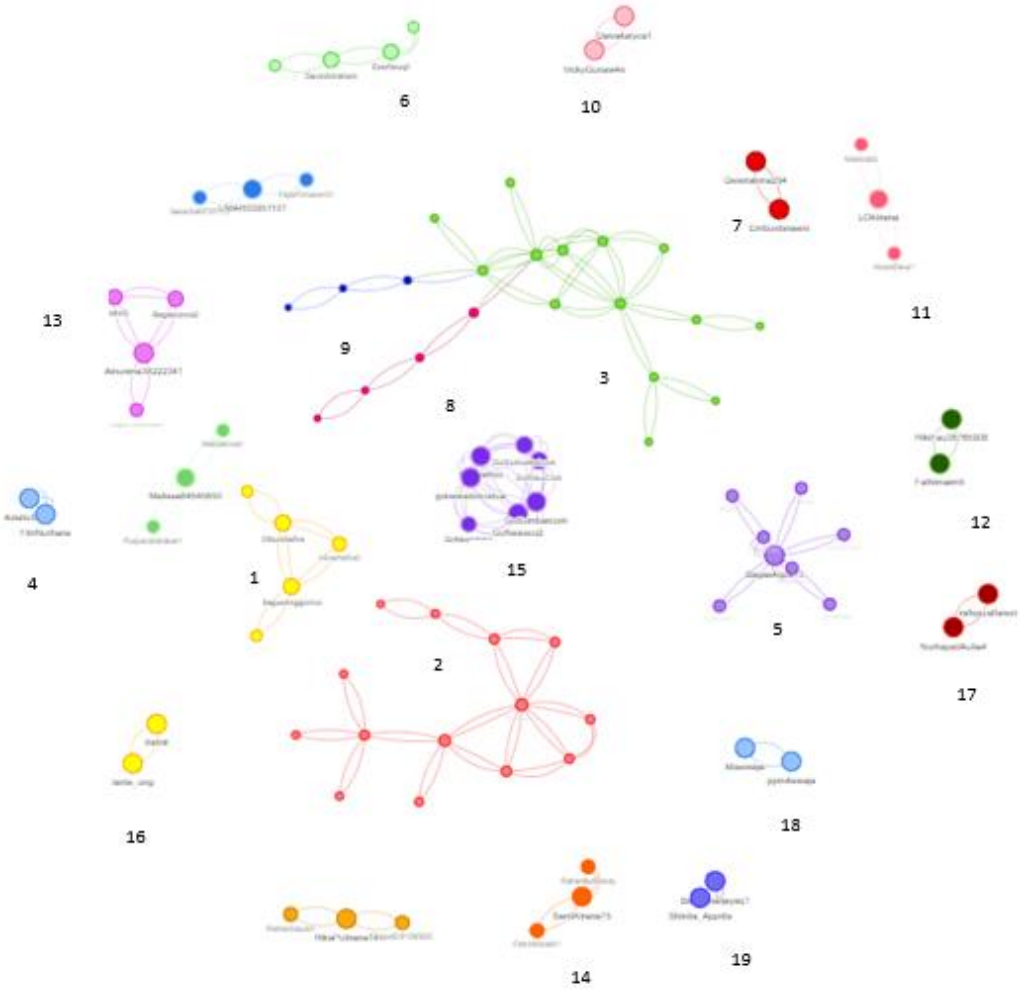

Figure 3. Visualization of all communities

The following is an analysis and visualization in figure that identify the 5 largest groups of accounts on the Twitter account network:

Community 1 consists of 5 nodes. These accounts are a collection of groups that discuss politics but have tweeted using a hashtag related to FPI. The node as the center of the community has a centrality value of 0.8 and the farthest node from the center has a centrality value of 0.5. All nodes in community 1 can be seen in the Figure 4 and Table 5.

Table 5. Result of closeness centrality node community 1

User	Centrality
BagusAnggorro	0.8
CNurshafira	0.8
Alicezhafira2	0.666666
AdiraArji	0.5
Davinamutiara6	0.5

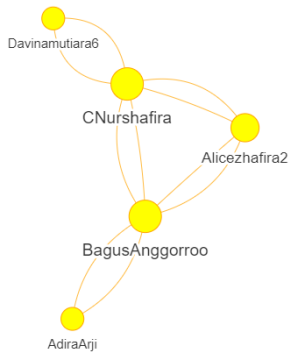

Figure 4. Visualization of community 1

 http://doi.org/10.26594/register.v6i1.1595

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

Community 15 consists of 7 nodes. These accounts are a collection of social media news groups that have discussed FPI. The node, as the center of the community, has a centrality value of 1 and the farthest node from the center has a centrality value of 0.75. The nodes having the same value means they also have the same number of relations or means it is resulted in the exactly same tweets. All nodes in community 15 can be seen in the Figure 5 and Table 6.

Table 6. Result of closeness centrality node Community 15

User	Centrality
GoAcehCo	1.0
GoNewsco2	1.0
goSumbarcom	1.0
gonewsdotcodua	1.0
GoRiauCom	0.8571428571428571
GoSumutdotcom	0.8571428571428571
GoNewsdotco	0.75

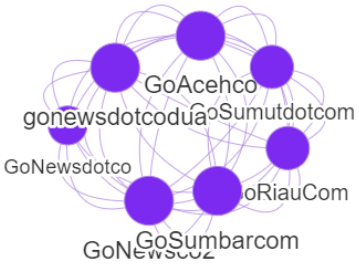

Figure 5. Visualization Community 15

Community 5 consists of 7 nodes. These accounts are a collection of groups supporting presidential candidates 02 but using a hashtag related to FPI. This Twitter accounts tweet the same content repeatedly, for example in BagasAgus13 account tweet "Allahu Akbar" 5 times. The node as the center of the community has a centrality value of 1 and the farthest node from the center has a centrality value of 0.53, the node that has the same value means that it has the same number of relations or because are really the same. All nodes in Community 5 can be seen in the Figure 6 and Table 7.

Table 7. Result of closeness centrality node Community 5

User	Centrality
BagasAgus13	1.0
GiaUmmu	0.5384615384615384
RYadie13	0.5384615384615384
Roby04185880	0.5384615384615384
Aiswan010	0.5384615384615384
miswantoaji	0.5384615384615384
Peju4ng	0.5384615384615384

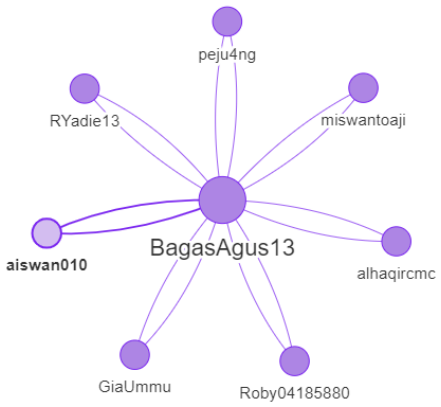

Figure 6. Visualization of Community 5

Community 3 consists of 14 nodes. These accounts are groups that attack presidential candidates 02 and have used hashtags related to FPI. In this community there is a section where nodes are not considered to be part of the community, because compared with the community center, tweet pairs from these accounts obtained less than 5. For example, if the last 10 tweets of the mimamananda2

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

account compared to the jessicaarina3 account there are 5 pairs of the same tweet, but when the hainessariati1 account is compared to the mimananda2 account the same tweet is less than 5 pairs, so the jessicaarina3 and hainessariati1 accounts are considered to be different communities. The node as the center of the community has a centrality value of 0.54 and the farthest node from the center has a centrality value of 0.24. The node that has the same value means that it has the same number of relations or because their tweets are really the same. All nodes in community 3 can be seen in the Figure 7 and Table 8.

Table 8. Result of closeness centrality node Community 3

User	Centrality
imamananda2	0.46511627906976744
fathulariz	0.43478260869565216
FarhanMubina3	0.4166666666666667
Hayachedva1	0.39215686274509803
PutriRahhyuu	0.39215686274509803
haryaharshita	0.37037037037037035
ShaniaJulia4	0.32786885245901637
GeraldRalph4	0.3225806451612903
AzahriSyifa	0.31746031746031744
AnnaVirgiana	0.3125
saskaraardhani	0.2857142857142857
Agraroberto8	0.25

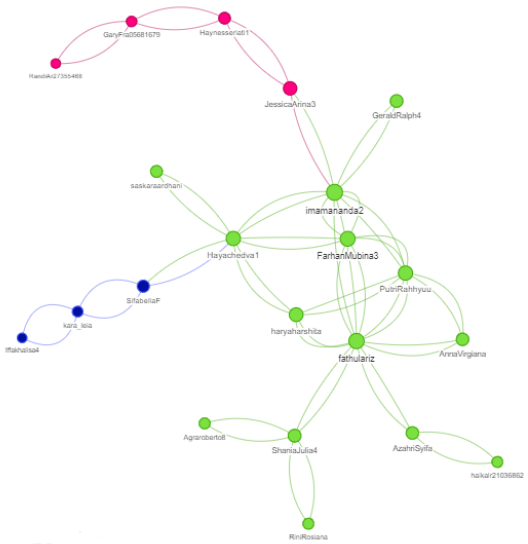

Figure 7. Visualization of *community* 3

Table 9. Result of closeness centrality node Community 2

User	Centrality
FitriaAuuliaa	0.4166666666666666
DoniFebri4n	0.52
GavinArfan1	0.4482758620689655
AugustaDeolinda	0.41935483870967744
Haidarlanden	0.41935483870967744
FarhanHilmawan5	0.3939393939393939
ArvinMaula4na	0.38235294117637056
Daivaanalise1	0.35135135135153157
ElenoAbraham	0.317073170731073
DewiAnggaraaai	0.3023255913953488
FirtiRatnaasarr	0.3023255913953488
SemiraDewi	0.3023255913953488
Deanakalista3	0.24528301886792453

Community 2 consists of 14 nodes frim accounts that attack the pair of presidential candidates 02 and have used hashtags related to FPI. Twitter accounts in this community tweeted about people power

ISSN 2502-3357 (Online)
A. Irsyad & N. A. Rakhmawati./Register 6 (1) 2020 pp. 22-31 ISSN 2503-0477 (Print)

and FPI chairman Rizieq Shihab. The node which is the center of the community has a centrality value of 0.54 and the farthest node from the center has a centrality value of 0.24. The node that has the same value means that it has the same number of relations or because the tweets are really the same. All nodes in community 2 can be seen in the Figure 8 and Table 9.

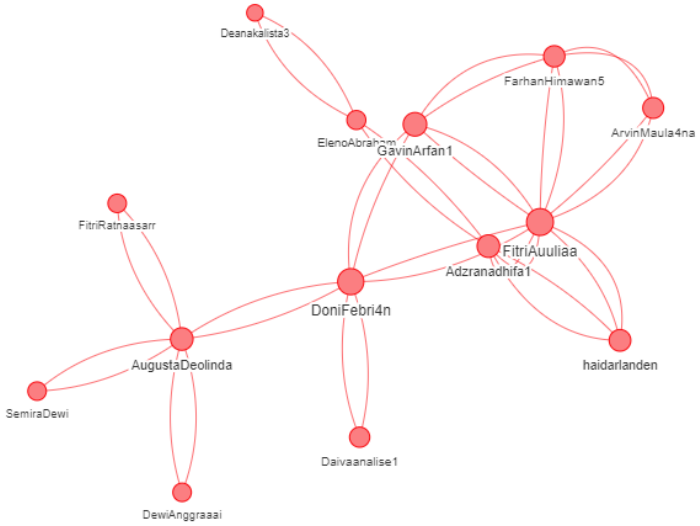

Figure 8. Visualization of Community 2

It can be seen from the results of the study, this research succeeded in detecting communities formed in social media in certain events. There are still few studies that study and presented communities the result in graphical form, such as research conducted by Fócil- Arias et al. [9] and Conover et al. [10] that did not use graph to display the results making it difficult to understand the shape and pattern of the community formed. In Indonesia, there has not been any research on the detection of social media account communities.

5. Conclusions

Based on the results, it can be seen that Louvain algorithm can be used to detect communities on Twitter. A high similarity value can be obtained by implementing cosine similarity algorithm. However, the cosine similarity does not pay attention to the order of the characters of each tweet resulting tweets with the same characters but different contents will be considered to have a high similarity. In communities clustering process, the level of similarity in tweets gives big impact of communities formation. Detection of communities on social media can be done using the Louvain algorithm.

6. References

[1] N. R. Fatahillah, P. Suryati and C. Haryawan, "Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech," in *International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia, 2017.

[2] I. A. Nur, M. A. Bijaksana and E. Darwiyanto, "Community Detection Menggunakan Genetic Algorithm dalam Social Network Twitter," in *eProceedings of Engineering*, 2015.

[3] Y. Zhang, Y. Wu and Q. Yang, "Community Discovery in Twitter Based on User Interests," *Journal of Computational Information Systems,* vol. 8, no. 3, p. 991–1000, 2012.

[4] C. N. Utami, W. Maharani and A. Adiwijaya, "Analisis dan Implementasi Community Detection Menggunakan Algoritma Girvan and Newman Dalam Sosial Network," Telkom University, Bandung, 2013.

[5] A. Riyani, M. Z. Naf'an and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *Jurnal Linguistik Komputasional (JLK),* vol. 2, no. 1, pp. 23-27, 2019.

[6] M. R. R. Gunaedi, I. Atastina and A. Herdiani, "Analisis dan Implementasi Algoritma Dynamicnet pada Deteksi Evolusi Komunitas di Media Sosial Twitter," in *e-Proceeding of Engineering*, 2018.

[7] D. R. Lazuardi, "Analisis Sentimen untuk Mengetahui Persepsi Kualitas Merek Menggunakan Text Mining dan Social Network Analysis Pada Konten Percakapan Di Media Sosial Twitter," in *eProceedings of Management*, 2014.

[8] S. Dutta, S. Ghatak, M. Roy, S. Ghosh and A. K. Das, "A graph based clustering technique for tweet summarization," in *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, India, 2015.

[9] C. Fócil-Arias, J. Zúñiga, G. Sidorov, I. Batyrshin and A. Gelbukh, in *Conference and Labs of the Evaluation Forum*, Dublin, Ireland, 2017.

[10] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonc¸alves, A. Flammini and F. Menczer, "Political Polarization on Twitter," in *Proceedings of The Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, Barcelona, Spain, 2011.

[11] F. P. Azali, "Klasifikasi Pengaduan Masyarakat Berbasis SMS dengan Metode Naive Bayes Classifier," Universitas Gadjah Mada, Yogyakarta, 2016.

[12] T. Arif, "Prediksi Perpindahan Pelanggan Industri Telekomunikasi Seluler Menggunakan Klasifikasi Sentimen Pada Situs Jejaring Sosial Twitter Menggunakan Support Vector Machine," Institut Teknologi Sepuluh Nopember, Surabaya, 2016.

[13] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing & Management,* vol. 50, no. 1, pp. 104-112, 2014.

[14] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 6, no. 4, pp. 1-33, 2007.

[15] Z. Pratama, E. Utami and M. R. Arief, "Analisa Perbandingan Jenis N-Gram Dalam Penentuan Similarity Text pada Deteksi Plagiat," *Citec Journal,* vol. 4, no. 4, pp. 254-263, 2017.

[16] O. Nurdiana, J. Jumadi and D. Nursantika, "Perbandingan Metode Cosine Similarity dengan Metode Jaccard Similarity pada Aplikasi Pencarian Terjemah Al-Qur'an dalam Bahasa Indonesia," *JOIN,* vol. 1, no. 1, pp. 59-63, 2016.

[17] M. Needham and A. E. Hodler, A Comprehensive Guide to Graph Algorithms in Neo4j, Neo4j, 2018.

[18] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* 2008.

[19] B. Wei and Y. Deng, "A cluster-growing dimension of complex networks: From the view of node closeness centrality," *Physica A: Statistical Mechanics and its Applications,* vol. 522, no. 15 May 2019, pp. 80-87, 2019.