Bit-Tech (Binary Digital - Technology)

Vol.6, No.1, August 2023
Available online at: http://jurnal.kdi.or.id/index.php/bt

Clustering Mental Health on Instagram Users Using K-Means Algorithm

Yuliastati Putri Sugiarta Karlim[1], Aditiya Hermawan[2]*, Ardiane Rossi Kurniawan Maranto[3]

[1)2)3)]Universitas Buddhi Dharma
Jl. Imam Bonjol No.41, Tangerang, Indonesia

[1)]yuliastati.putris@ubd.ac.id

[2)]aditiya.hermawan@ubd.ac.id

[3)]ardiane.rossi@ubd.ac.id

Article history:

Received 15 July 2023;
Revised 31 July 2023;
Accepted 14 August 2023;
Available online 25 August 2023

Keywords:

Clustering
Elbow Method
Instagram
K-Means
Mental Health

Abstract

The use of Instagram too often can have an impact on the mental health of its users. Mental health that is not good requires early treatment so that it does not have a widespread impact on other health. Mental illness requires a professional to treat it as an effort to prevent a disease from getting worse. However, the stigma attached to sufferers is one of the significant causes behind the reluctance to seek treatment. Therefore we need a way so that Instagram users can find out for themselves the condition of their mental health. One way is to do Clustering the use of Instagram so that it can provide an early indication of a person's mental health. From the proposed model we can find out the categories of 600 respondents who were collected using a questionnaire with 10 main attributes. The proposed model is k-means with 3 clusters determined using the elbow method. In this study, the last centroid obtained through calculations was used to evaluate the k-means by comparing the results of the k-means calculations with the results of psychologists. The results of the K-means evaluation have an accuracy of 73.83% so that the last centroid can be applied to web-based applications that have been created. This mental health clustering model is expected to be able to help the community to get mental health conditions early and reduce the negative stigma that exists and can be used as evaluation material in using social media more wisely.

I. INTRODUCTION

Humans interact with one another through face to face or by using various kinds of social media. Nowadays, social media has become a necessity that cannot be separated from someone. There is social media that is a means of communication to social media that functions as a virtual world gallery. One of the most widely used social media right now is Instagram. According to data released by Hootsuite and We Are Social as of January 2020, active users of Instagram worldwide have reached 1,221,000,000 users [1].

Instagram is a social media platform based on photos and videos. There are many features launched by Instagram, from filters, stickers, locations to polling systems. The reason someone uses Instagram is as a means to express themselves, and to meet the needs of caring and caring for others. This reasoning is by Maslow's theory of needs, namely love, belonging, and self-esteem where there will be a feeling of worth and trust in one's strengths and abilities, if others are paying attention.

Over time, Instagram began to hurt its users. Someone who uses Instagram more than 2 hours a day and 58 times a week is more susceptible to adverse effects [2]. Adverse effects that arise such as anxiety, depression, loneliness, lack of sleep, intimidation, self-image disorders, and fear of missing information about others. These adverse effects make a person mentally unhealthy. Therefore, Instagram has been said to be the most social media that hurts the mental health of its users [3].

Mental health or what is commonly referred to as mental health is a condition where a person can realize his potential, can cope with stress, can work productively and can make a good contribution to society or others [4]. To determine one's mental health position, it cannot be seen in black and white but must look at existing indicators. If

* Corresponding author

doi: 10.32877/bt.v6i1.880

someone is mentally unwell or has a mental disorder, a psychologist or psychiatrist can help someone's mental recovery. But the stigma that exists in Indonesian society, in general, is that mental health is not important, and going to professionals such as psychologists or psychiatrists indicates that the person is crazy and just throwing money away, even though mental health is as important as physical health. Moreover, Indonesia is the fourth country with the most Instagram users, amounting to 94,3 million active users in October 2021 [5]. Then a classification system must be made to provide an initial indication of a person's mental health based on Instagram usage using the K-Means algorithm. The K-Means algorithm is a popular clustering algorithm because it can only process data with numeric attributes and determine the number of clusters to be formed from the start.

II. Literature Review

A. K-Means Algorithm

K-Means is one of the most widely used clustering algorithms because it is easy to implement. The K-Means algorithm is a distance-based algorithm that attempts to partition data into clusters. The K-Means algorithm has the following procedures [6].

a. Determine the number of clusters.
b. Randomly allocate data into clusters.
c. Calculate the center of the cluster (centroid) or the average of the data that is in each cluster :

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \tag{1}$$

Where M is amount of data in a cluster, i is i-feature in a cluster, and p is data dimension. The formula is done as many as p dimensions, so that i start from 1 to p and to measure the distance to centroid data will use the Euclidean Distance formula as follows :

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p}\left|x_{2j} - x_{1j}\right|^2} \tag{2}$$

Where D is distance between data x2 and x1 and |.| is an absolute value, p is data dimension, x2j is coordinates of object i on dimension k, and x1j is coordinates of object j on the dimension k.

d. Allocate each data to the closest centroid / average based on the comparison of distances between centroid data and each cluster.:

$$a_{il} = \begin{cases}1 \\ 0\end{cases} d = min\{D(x_i, C_l)\} \tag{3}$$

Where ail is the membership value of point xi to the center of the C1 cluster, d is the shortest distance from data xi to K cluster after comparison, and C1 is the 1st cluster center (centroid).

e. Repeat steps (c) and (d), until the cluster center results are not changed.

B. Elbow Method

Elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of the results of the comparison between the number of clusters that will form an elbow at a point [7]. This method provides an idea by selecting the number of clusters and then increasing the number of clusters to be used as a data model in determining the best cluster. The percentage of the calculation produced becomes a comparison between the number of clusters added [8].To get a comparison of cluster values is to calculate the SSE (Sum of Square Error) of each cluster in the following equation [9] :

$$SSE = \sum_{K=1}^{K}\sum_{i\in S_k}\|y_i - C_k\|^2 \tag{4}$$

Where K is the number of clusters, Yi is the i data, and Ck is the average value of the K cluster. After calculating, there will be some K values that have decreased the most, and then the K value will decrease slowly until the result of the K value is stable. For example, in Figure 1 there is a drastic decrease and form an elbow at point K = 3, then the ideal K cluster value is K = 3.

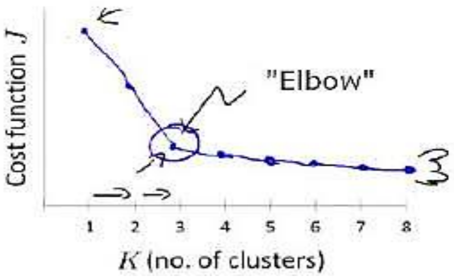

Fig. 1 Elbow graph method

C. Knowledge Discovery Database (KDD)

Data mining is the core of Knowledge Discovery Databases (KDD) where algorithms explore data, build models, and discover unknown patterns [10]. KDD is solving the problem by analyzing the data in the database. KDD has the following stages which are described in Figure 2 [11] :

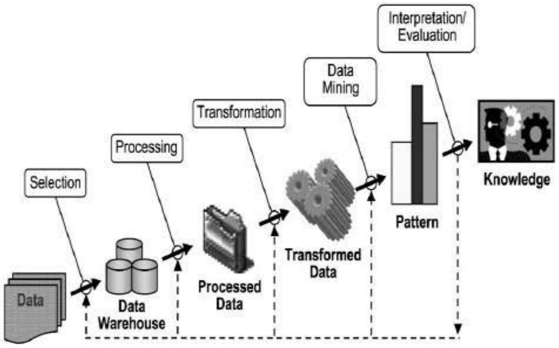

Fig. 2 Stages of the KDD process

a. Data Selection
The selection of data needs to be done before the information gathering stage in KDD begins. The results of the data selection that will be used for the data mining process are stored in one file and separate from the operational database.

b. Pre-processing / Cleaning
This cleaning process includes removing data duplication, checking inconsistent data, correcting errors in the data, enriching or enriching existing data with relevant data that is needed in KDD.

c. Transformation
Coding is the process of transformation of data that has been selected so that the data is following the data mining process. This is a creative process and is very dependent on the type or pattern of information that will be sought in the database.

d. Data Mining
Data mining is the process of finding interesting patterns or information in selected data by using certain techniques or methods following the overall goals and processes of KDD.

e. Interpretation / Evaluation
This stage involves checking the information generated on the facts or hypotheses that already exist and displaying the results of information obtained from the data mining process in a form that is easily understood by the user.

D. Mental Health

Mental health is an individual who is free from all psychiatric symptoms or mental disorders with the realization of harmony between mental functions and have the ability to deal with problems that occur [12]. Mental health is the realization of true harmony between mental functions and the creation of self-adjustment between humans and themselves and their environment based on faith and piety and aims to achieve a meaningful and happy life in the world and the hereafter [13].

Someone who experiences mental-emotional disorders will experience a decline in function in the realm of family, work, education, community, and society [14]. There are several signs and symptoms of mental disorders, namely as follows [15]:

a. Cognitive disorders are disorders of mental processes that have a relationship that is realized and maintained by individuals with their environment. Mental processes include sensations and perceptions, attention, memory, associations, considerations, thoughts, and awareness.
b. Attention Disorders are disturbances in concentration and energy concentration.
c. Memory Disorders are disturbances of awareness signals and the ability to store, record, and produce content.
d. Disorders of Association are disturbances of the impression or description of memories caused by feelings, impressions, or images of memories in mental processes.
e. Disorders of Consideration are disorders of mental processes that provide consideration or assessment of the intent and purpose of the activity.
f. Mind disorders are part of one's knowledge.
g. Impaired consciousness is a disturbance in a person's ability to make a relationship or limitation between himself and the environment through his five senses.
h. Disruption of the Will is a disruption in the process of the desire carried out to achieve the goal after consideration and then decided.
i. Emotional and Affective Disorders, emotions are conscious experiences and affect bodily activities that produce kinetic and organic sensations. Affect is the life of a person's emotional feeling or tone, pleasant or not, that accompanies a thought, usually lasting for a long time and is rarely accompanied by physiological components.
j. Psychomotor disorders are body movements that are influenced by mental states

III. METHODS

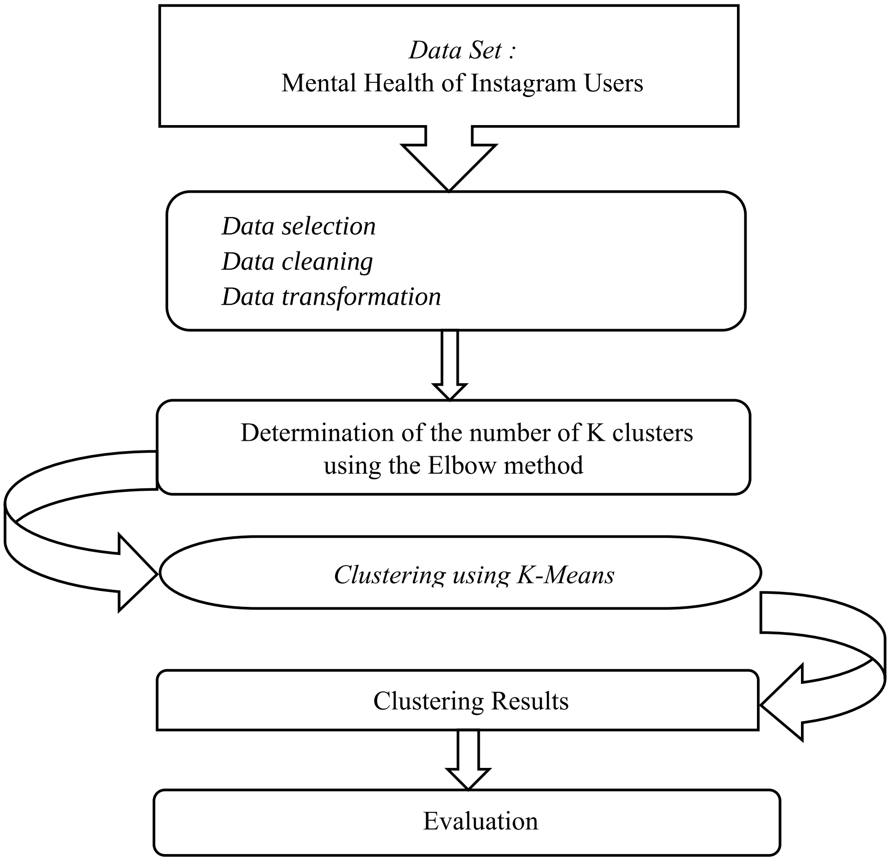
Data Set :
Mental Health of Instagram Users
Data selection
Data cleaning
Data transformation
Determination of the number of K clusters using the Elbow method
Clustering using K-Means
Clustering Results
Evaluation

Fig. 3 Clustering Method for Mental Health

IV. RESULTS

A. Data Selection

The data used are primary data obtained using a questionnaire. The questionnaire was aimed at 600 Instagram users in Indonesia. In obtaining these data, the authors make a selection or selection of data from a collection of data obtained from the questionnaire and carried out before the stage of extracting information. In the selection of these data, the selection of attributes that will be used in this study has been done. Selection of attributes based on previous research [16]. The attributes used are shown in Table 1.

TABLE 1
Attributes Detail

Number	Attributes	Description
1	Acc	Number of Instagram accounts
2	Freq	Frequency accessing Instagram in a day
3	Duration	Duration accessing Instagram in a day
4	Addiction	Addicted to access
5	Edit	photo/video editing
6	Like	Signs like photos/videos
7	Com	Comments
8	Envy	Envy with other people's photos/videos
9	FOMO	Fear of missing information
10	Foll	Followers account

B. Data Cleaning

In this step, a process of data elimination for entries lacking values, removal of duplicates, and consistency checks is conducted. Additionally, data errors are rectified, and irrelevant attributes for this research are removed. In Table 2 below the results after the data set is selected

TABLE 2
Dataset after selection

Number	Acc	Freq	Duration	Addiction	Edit	Like	Com	Envy	FOMO	Foll
1	1	Less than 8 times/day	Less than 1 hour/day	2	1	4	1	1	1	5
2	2	8-16 times/day	More than 2 hour/day	4	5	3	2	1	2	4
3	1	8-16 times/day	More than 2 hour/day	3	1	1	1	1	1	1
4	1	Less than 8 times/day	Less than 1 hour/day	2	1	1	1	2	1	1
5	2	More than 16 times/day	More than 2 hour/day	5	2	3	1	1	3	5
6	1	8-16 times/day	Less than 1 hour/day	3	3	3	2	2	2	4
7	2	More than 16 times/day	Less than 1 hour/day	4	4	2	2	1	2	5
8	1	8-16 times/day	More than 2 hour/day	5	5	3	4	3	4	4
9	1	Less than 8 times/day	Less than 1 hour/day	3	3	3	4	4	4	3
10	3	8-16 times/day	1-2 hour/day	5	5	5	5	1	5	5

Data Transformation

During this phase, discretization is conducted to conform to the desired analysis and the required data structure, simplifying the data mining procedure. This process of discretization is applied to 2 attributes and the results can be seen in Table 3.

1. Frequency:
   a. Less than 8 times/day = 1
   b. 8-16 times/day = 2
   c. more than 16 times/day = 3
2. Duration:
   a. Less than 1 hour/day = 1
   b. 1-2 hour/day = 2
   c. More than 2 hour/day = 3

TABLE 3
Data after transformation

No	Acc	Freq	Durasi	Candu	Edit	Like	Com	Iri	FOMO	Foll
1	1	1	1	2	1	4	1	1	1	5
2	2	2	3	4	5	3	2	1	2	4
3	1	2	3	3	1	1	1	1	1	1
4	1	1	1	2	1	1	1	2	1	1
5	2	3	3	5	2	3	1	1	3	5
6	1	2	1	3	3	3	2	2	2	4
7	2	3	1	4	4	2	2	1	2	5
8	1	2	3	5	5	3	4	3	4	4
9	1	1	1	3	3	3	4	4	4	3
10	3	2	2	5	5	5	5	1	5	5

Determination The Number of Clusters

Some K values have decreased the most and then the K value will decrease slowly until the result of the K value is stable. Based on the results of the SSE (Sum of Square Error) calculation which is shown in Table 2 with as many as 600 data, there is an SSE value that has decreased dramatically and formed an elbow on the graph at point K = 3, then the ideal K cluster value is the number of clusters as much as 3. It is not too far away and is the best number of clusters compared to other clusters.

TABLE 4
SSE results for each cluster

Cluster	SSE	Gap
1	981.339	981.339
2	673.7	307.639
3	448.989	224.711
4	424.523	24.466
5	389.96	34.563
6	365.45	24.51

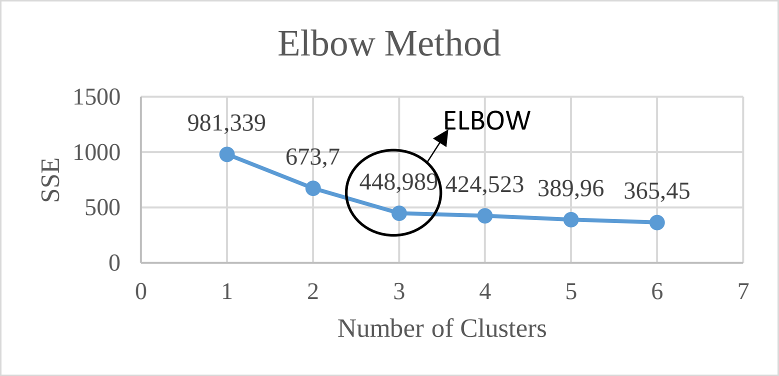

Fig. 4 SSE graph

Based on the results of the Elbow method in Figure 4 which was carried out on the collected data set, the clusters to be made are 3 Clusters. The 3 clusters to be searched are labeled as follows namely C1 for the 'Healthy' category, C2 for the 'Worrying' category, and C3 for the 'Very Worrying' category.

Cluster with K-Means

To find clusters from a dataset of 600 data that will be divided into 3 clusters according to the Elbow Method for the best number of clusters from the existing dataset. Looking for clusters from existing data using the K-Means Algorithm to find the proximity of each data to the centroid value of each cluster.

TABLE 5
Centroid of the average value of each cluster

Centroid	Acc	Freq	Durasi	Candu	Edit	Like	Com	Iri	Fomo	Foll
C1	1.2684	1.6623	1.5152	2.6147	2.3853	1.5628	1.7619	1.3377	1.4719	2.8225
C2	1.2304	2.1832	1.7644	3.0000	2.8534	1.7435	2.6545	3.1675	3.6387	3.5288
C3	1.7472	2.3427	2.4045	3.8315	3.8034	3.4270	3.2135	2.5955	2.9045	4.1292

From the centroid value obtained in Table 3, every data in the dataset is searched for the proximity of the data to the centroid value of each cluster. From this data, the following clusters were obtained: Cluster 1 for the mental health category in the 'Healthy' level with 231 data, cluster 2 for the mental health category in the 'Worrying' level with 191 data and cluster 3 for the mental health category in the 'Very Worrying' level, which has 178 data.

K-Means Evaluation

In this process, an examination of the results of the information on the facts or hypotheses will be carried out by comparing the results of the clustering evaluation with the results that have been carried out by psychologists. The following is the comparison in Table 4:

TABLE 6
Comparison of the results of Psychologists and K-Means clustering

Number	Psychologist evaluation results	Clustering Results (K-Means)	Information
1	Worrying	Very Worrying	FALSE
2	Worrying	Very Worrying	FALSE
3	Healthy	Healthy	TRUE
4	Healthy	Healthy	TRUE
5	Healthy	Healthy	TRUE
6	Very Worrying	Very Worrying	TRUE
7	Very Worrying	Very Worrying	TRUE
8	Healthy	Healthy	TRUE
9	Healthy	Healthy	TRUE
10	Worrying	Very Worrying	FALSE
11	Healthy	Healthy	TRUE
12	Worrying	Healthy	FALSE
13	Worrying	Very Worrying	FALSE
14	Very Worrying	Very Worrying	TRUE
15	Worrying	Very Worrying	FALSE
16	Very Worrying	Worrying	FALSE
17	Very Worrying	Very Worrying	TRUE
18	Healthy	Healthy	TRUE
19	Worrying	Very Worrying	FALSE
20	Worrying	Worrying	TRUE

Based on table 4, there are data from 600 different data between the results of evaluations conducted by psychologists and the results of clustering using K-Means. To calculate accuracy will be done using the following equation :

$$\text{System Accuracy} = \frac{\text{number of correct results}}{\text{sum of all data}} \text{x}100\% \quad (5)$$

$$= \frac{600-157}{600} \text{x}100\% = 73.83\%$$

The accuracy of the test results showed that the system performance was good by achieving results of 73.83%.

V. Discussion

The Clustering model for the mental health of Instagram users still needs some improvements in order to perform Clustering more accurately. The more datasets and the more diverse respondents in use, of course, the better the clustering that is formed. In addition, the evaluation of clustering results that need to be checked with the help of more mental health workers will make the clustering accuracy value made more valid. You can also compare the results of the K-Means Clustering with the Clustering Algorithm to see the difference in the results so that the clustering results can help Instagram users as an early indication to check their mental health. This is done to assist social media users, especially Instagram users, in becoming wiser in their usage to safeguard their mental well-being. It can also invite other people to care more about their mental health conditions.

VI. CONCLUSIONS

Based on the making of a new model with the K-Means algorithm that has been made and explained in this study. This study can cluster mental health data based on the use of Instagram using K-Means with an evaluation value of 73.83%. The formed K-Means model can be used to determine mental health data categories. There are 3 categories of mental health clusters, namely Healthy, Worrying, and Very Worrying clusters. For further research is expected to use a larger number of datasets and data obtained in real time for better clustering results and can provide more detailed indications of mental health such as anxiety, depression, narcissism. You can also add sentiment and text analysis to posts on social media so that they can provide additional, more specific information in diagnosing mental health when using social media.

REFERENCES

[1] S. Kemp, "Digital 2021 Global Overview Report," *In Global Digital Insights*, 2021.

[2] A. Macmillan, *Instagram Is the Worst Social Media for Mental Health.* 2017.

[3] Royal Society For Public Health, *Status Of Mind : Social media and young people's mental health and wellbeing.* 2017.

[4] World Health Organization, "Mental health: strengthening our response," 2018.

[5] Statista, *Leading Countries Based on Instagram Audience Size as of October 2021*. 2021.

[6] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Andi, 2013.

[7] T. Soni Madhulatha, "AN OVERVIEW ON CLUSTERING METHODS," *IOSR J. Eng.*, vol. 02, no. 04, pp. 719–725, Apr. 2012, doi: 10.9790/3021-0204719725.

[8] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 975–8887, 2014, doi: 10.5120/18405-9674.

[9] T. . Kodinariya and P. . Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.

[10] A. Chailes, A. Hermawan, and D. Kurnaedi, "Penerapan Metode Data Mining Untuk Menentukan Pola Pembelian Dengan Menggunakan Algoritma," *J. Algor*, vol. 1, no. 2, pp. 1–8, 2020.

[11] R. T. Vulandari, "Pengelompokan Tingkat Keamanan Wilayah Jawa Tengah Berdasarkan Indeks Kejahatan Dan Jumlah Pos Keamanan Dengan Metode Klastering K-Means," *J. Ilm. SINUS*, vol. Vol 14, No, no. ISSN :1693-1173, pp. 59–72, 2016, doi: http://dx.doi.org/10.30646/sinus.v14i2.252.

[12] B. Bukhori, "Hubungan Kebermaknaan Hidup Dan Dukungan Sosial Keluarga Dengan Kesehatan Mental Narapidana (Studi Kasus Nara Pidana Kota Semarang)," *Addin*, vol. 4, no. 1, pp. 1–19, 2012.

[13] D. V. Fakhriyani, *Kesehatan Mental*. Pamekasan: Duta Media Publishing, 2019.

[14] Y. Kurniawan and I. Sulistyarini, "Komunitas Sehati (Sehat Jiwa dan Hati) Sebagai Intervensi Kesehatan Mental Berbasis Masyarakat," *Insa. J. Psikol. dan Kesehat. Ment.*, vol. 1, no. 2, p. 112, 2016, doi: 10.20473/jpkm.v1i22016.112-124.

[15] A. Nasir and A. Muhith, *Dasar-Dasar Keperawatan Jiwa : Pengantar dan Teori*. Jakarta: Salemba Medika, 2011.

[16] R. Mondoano, Nuraeni Aprilia; Mayasar and F. Gunawan, "Instagram And Mental Health Of The Students In Islamic Higher Education Of Southeast Sulawesi Indonesia," 2018, no. September, pp. 26–34.