# ANALYSIS OF FACEBOOK USER PROFILING USING CLUSTERING IMPLEMENTATION

## A Case Study : Facebook User Profiling

[1)]Fauzan Atha Prakoso, [2)]Noer Alam Yahya, [3)] Dhian Satria Yudha Kartika
[1,2,3] Information Systems, Faculty of Computer Science
East Java "Veteran" Pembangunan Negeri University
Surabaya, East Java, Indonesia
[1)]19082010080@student.upnjatim.ac.id
[2)]19082010090@student.upnjatim.ac.id
[3)] dhian.satria@upnjatim.ac.id

***Abstract*—** In this era that is increasingly developing, it is possible that nearly some human beings use social media as a means of verbal exchange between users. It is not only a means of communique but using social media is used to show their own daily activities or habits. Social media is a media that functions as a medium for socializing from individual to individual based online using digital technology. Where the communication process is done via the internet. Of the several media used for socializing, Facebook is a very popular application. In the use of Facebook which can be accessed from all devices with the internet being the main role, there is a like feature provided by Facebook as a feature for socializing, of course there will be a comparison between the preferred data from some of these devices and from these data we can wonder why people chose mobile over computer or pc to access facebook. To calculate and analyze the existing data, we compared the data by conducting research and looking for Data Mining with the Clustering method using the K-Means Clustering algorithm.

***Keywords*—*Social media, Facebook, Clustering, Data mining, K-Means.***

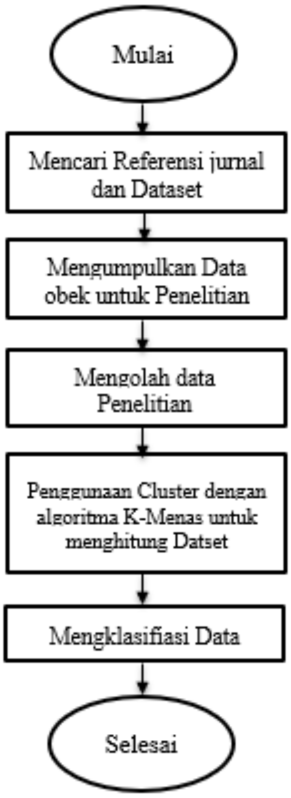
## I. INTRODUCTION

The use of social media is increasing considering the current era The use of social media has played a role in everyday life for several years final. The use of social media as a method of communication between users was created to make it easier for the general public. one of the most popular social media used in parts of the world and in Indonesia is the social media Facebook. The platform is used by more than hundreds of millions of people. Popularity on social media Facebook has become very well known to the wider community. That's why the Facebook crowd can This is because social media like Facebook is able to connect hundreds of millions of people people from all over the world without knowing the boundaries of geographic value [1]. The social media platform Facebook has become a platform favored by the general public because of the ease of long-distance communication that is easy to use. Facebook is a social media platform that plays a role for effective digital communication make it easier for users to interact with each other without having to meet in person. Social media such as Facebook can recap some data that are mutually exclusive connect using account profiles and personal interactions between users or through multiple groups, as well as content that is liked or shared. Data set can be obtained by using existing software to obtain (scrapting) Facebook information. It's just that large data sets are hard to find because Facebook has implemented some privacy settings on the news users, so that the news obtained is only limited to data originating from users who have friendships with scrapting users [2].

One of the methods in records data mining is Clustering, and this method is one of the most valid to be used to solve various kinds of complex problems in science computers and statistics [3]. Clustering is a technique derived from data mining which aims to group statistics based on the characteristics of similarity between one fact with other records [4]. One of the clustering algorithms that can be used in grouping data according to similarity characteristics is K-Means Clustering. On data Clustering using the K-Means Algorithm has an algorithm using iterative grouping that partitions the data set into multiple clusters already affected at the beginning which can be implemented and executed using quite fast and practical to follow the situation [5]. Research on the K-Means algoritma algorithm carried out by researchers to help the needs of various fields one is in sales or business. K-Means algorithm that our group used in this study because its implementation is quite easy and time consuming needed in carrying out the process does not take long and is easy to adapted and studied, then this research on Facebook data clustering hopefully it can help in the clustering process the number of likes you get via the web with mobile by the user

## II. METHODOLOGY

In this research, several stages are used in conducting research from looking for a refrence journals or references, looking for a Dataset reference to do a several stages such a determining the objeect data and use of the K-Means algorithm to cluster. here I use the clustering methodology using K-means to profile data from Facebook in the form of how many likes there are on mobile_likes and www_likes which is on the user profiling database from the facebook data that has been provided. From several k-means clustering methods, data will be generated in the form of an explanation generated from the Google Collabs source code. and this is the flow of the research method that I use with an explanation in the picture below.


Mulai
Mencari Referensi jurnal dan Dataset
Mengumpulkan Data obek untuk Penelitian
Mengolah data Penelitian
Penggunaan Cluster dengan algoritma K-Menas untuk menghitung Datset
Mengklasifiasi Data
Selesai

. 1. PROCCESS OF DATA PROCESSING

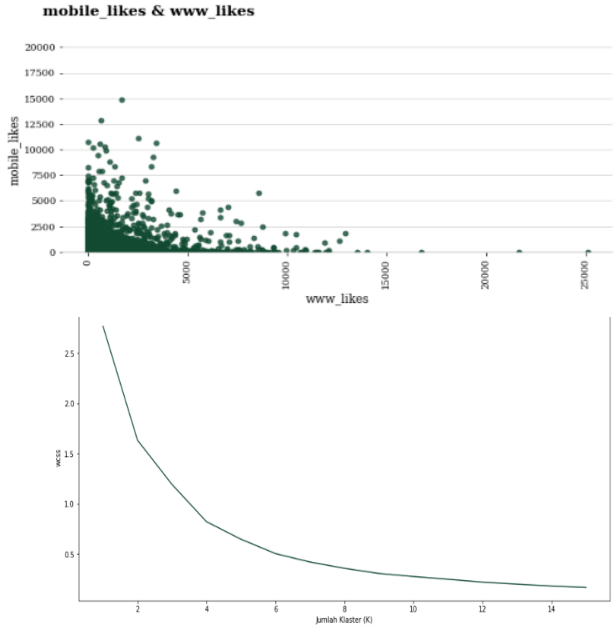
The explanation of the stages of implementing the K-Means algorithm that we did is as follows :

1. Looking for journal references and datasets to meet research needs.
2. Collecting object data on several types of datasets that will be taken for each data have each data type such as user id, age, gender, friend, mobile likes, www likes, like received.
3. Process the data and determine how many clusters (k) of the total data from the dataset that has been collected then chooses the centroid which is generally done random.
4. Clustering calculations using the K-Means cluster algorithm from the dataset profiling facebook (Mobile likes & www likes) by doing calculations using calculations from existing datasets using the centroid which has been specified. The formula used in the calculation of the distance as following.
5. Classify data by grouping objects. To define Cluster is to determine what object will be

used as a Clustering benchmark, here we define 3 Clusters.

## III. Result and Discussion

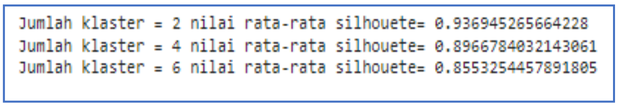
1) Showing data of mobile likes and www likes in Graphical Form.



2. GRAPHIC OF CAMPARISON BETWEEN WWW LIKES AND MOBILE LIKES

2) Determine Silhouette

The number of Clusters and Silhouettes, the largest is 1 Cluster, however we decided to use 3 Clusters, because of the amount of data in order to be able to meet suitability objective. Like a Very Many, Average, Vey Less



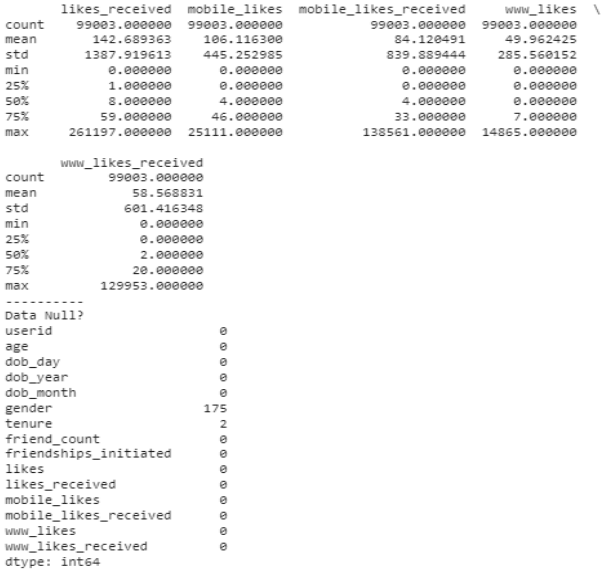
```
Jumlah klaster = 2 nilai rata-rata silhouete= 0.936945265664228
Jumlah klaster = 4 nilai rata-rata silhouete= 0.8966784032143061
Jumlah klaster = 6 nilai rata-rata silhouete= 0.8553254457891805
```

3. CLASS OF CLUSTERING

3) Implementation K-Means Clustering

a) Input Data

In this study, several types of data were input to be taken for each data each has its own data type such as user id, age, gender, friend, mobile likes, www likes, likes received.



```
       likes_received  mobile_likes  mobile_likes_received     www_likes  \
count    99003.000000  99003.000000           99003.000000  99003.000000
mean       142.689363    106.116300              84.120491     49.962425
std       1387.919613    445.252985             839.889444    285.560152
min          0.000000      0.000000               0.000000      0.000000
25%          1.000000      0.000000               0.000000      0.000000
50%          8.000000      4.000000               4.000000      0.000000
75%         59.000000     46.000000              33.000000      7.000000
max     261197.000000  25111.000000          138561.000000  14865.000000

       www_likes_received
count        99003.000000
mean            58.568831
std            601.416348
min              0.000000
25%              0.000000
50%              2.000000
75%             20.000000
max         129953.000000
----------
Data Null?
userid                     0
age                        0
dob_day                    0
dob_year                   0
dob_month                  0
gender                   175
tenure                     2
friend_count               0
friendships_initiated      0
likes                      0
likes_received             0
mobile_likes               0
mobile_likes_received      0
www_likes                  0
www_likes_received         0
dtype: int64
```

4. CALCULATING MEANS, MIN, MAX ETC ON WWW_LIKES RECEIVED, WWW_LIKES, MOBILE_LIKES AND MOBILE_LIKES RECIEVED

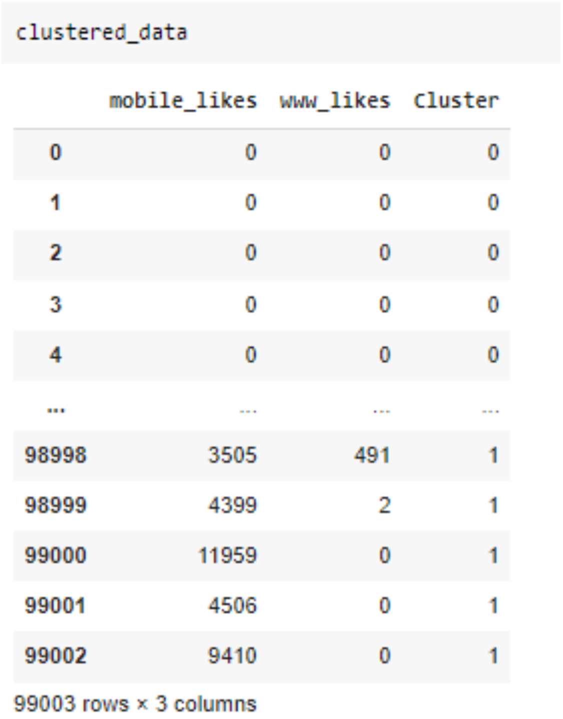
b) Determining the Number of Clusters and Centroid Value

The next stage is to determine the number of clusters or groups of initial centroid values, where the results of the formulation automatically show data results as follows.

| Centroid | 0 | 1 |
|---|---|---|
| Mobile_likes | 69,6 | 3097.3 |
| www_likes | 41,9 | 713.3 |
| Rata-Rata | 55,8 | 1.905,3 |

c) Define Group and Recalculate Centroid Value

At this stage, the system calculates the gap between the cluster center point and the point of each object (Facebook Data) which will then be grouped according to the comparison and selected through the closest distance between the data from the dataset and the cluster center, this distance indicates that the data was in a groups with the nearest Cluster center and compare the results of several clusters and later taken from the smallest.

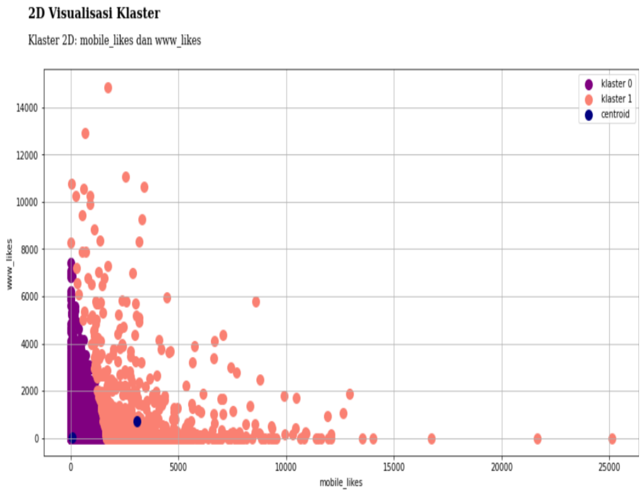


```
clustered_data
```

|  | mobile_likes | www_likes | Cluster |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| ... | ... | ... | ... |
| 98998 | 3505 | 491 | 1 |
| 98999 | 4399 | 2 | 1 |
| 99000 | 11959 | 0 | 1 |
| 99001 | 4506 | 0 | 1 |
| 99002 | 9410 | 0 | 1 |

99003 rows × 3 columns

5.CLUSTERED DATA OF MOBILE_LIKES AND WWW_LIKES



d) Grouping Results

After getting the results from the literacy process by K-Means Clustering, we get the results of grouping each Cluster are.

With Description of Each Cluster :

| Cluster | Kategori | Penggunaan Mobile_Likes |
|---|---|---|
| Cluster 0 | Kurang | 69.6 |
| Cluster 1 | Banyak | 3079.3 |

| Cluster | Kategori | Penggunaan www_Likes |
|---|---|---|
| Cluster 0 | Kurang | 41.9 |
| Cluster 1 | Banyak | 713.3 |

6. DEFINING EACH CLUSTER MEANS IF CLUSTER IS CALLED LESS (KURANG) WHILE CLUSTER 1 IS CALLED MANY (BANYAK)

## IV. CONCLUSIONS AND RECOMMENDATIONS

Based on the results obtained using the K-Means algorithm on 99004 facebook data based on www_likes and mobile_likes it produces two clusters, namely Less and Many. C0 (41.9, 69.6, 55.8), C1 (3079.3, 713.3, 1905.3). with the above results can it is concluded that there are more Mobile_likes than WWW_likes as the result above as well we can take a conclusion that people like to use mobile likes more due to their use of mobile phones and because its really easy to access while on web its bit confusing to access on mobile phone and nowdays people chose mobile over computer for day to day use. Suggestion from The research that we have concluded is by adding another research model and compared with some data from other platforms in order to get results in some other social media.

## REFERENCES

[1] Heidemann, J., Klier, M., & Probst, F. “Online social networks: A survey of a global phenomenon”. Computer Networks, 56(18), 3866– 3878, 2012.

[2] Rohman, Abdul, Ardani Yustriana Dewi, Kemas M. Irsan Riza, and Takdir. “Sosial Graf untuk

Visualisasi Data Facebok Menggunakan Visual Interaction System (Vis.js), 2014.

[3] D. Anggarwati, O. Nurdiawan, I. Ali and D. A. Kurnia, "Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri," JURNAL DATA SCIENCE & INFORMATIKA (JDSI), vol. Vol. 1 No. 2, pp. 58-62, 2021.

[4] J. Hutagalung and F. Sonata, "Penerapan Metode K-Means Untuk Menganalisis Minat Nasabah Asuransi," JURNAL MEDIA INFORMATIKA BUDIDARMA, Vols. Volume 5, Nomor 3, pp. 1187-1194, 2021.

[5] N. Mirantika, A. T. Ain and F. D. Agnia, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat," JURNAL NUANSA INFORMATIKA, vol. Volume 15 Nomor 2, pp. 92-98, 2021.

[6] Heidemann, J., Klier, M., & Probst, F. "Online social networks: A survey of a global phenomenon". Computer Networks, 56(18), 3866–3878 , 2012.

[7] Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. "On the evolution of user interaction in Facebook. Proceedings of the 2nd ACM Workshop on Online Social Networks" - WOSN '09, 37. 2009.

[8] Rohman, Abdul, Ardani Yustriana Dewi, Kemas M. Irsan Riza, and Takdir. "Sosial Graf untuk Visualisasi Data Facebok Menggunakan Visual Interaction System (Vis.js), 2014.

[9] Aggarwal, Charu C. Social Network Data Analytics. Springer, 2011.

[10] Setatama, M. S., & Tricahyono, D. Implementasi Social Network Analysis dalam Penyebaran Country Branding "Wonderful Indonesia." 2(2), 14, 2017.