CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

4.1.1 Data Collecting

Data obtained from Kaggle [3], the data which is divided into 8 columns contain 434.891 data. The data is in CSV format and then imported into Python Pandas.

Table 4. 1 Example Data

Date	Funny	Helpful	Hour Played	Is Early Access Review	Recommendation	Review	Title
2019-02-10	2	4	578	False	Recommended	&gt Played as German Reich&gt Declare War on Belgium&gt so go through France&gt…	Expansion – Hearts of Iron IV: Man the Guns
2019-02-10	0	0	184	False	Recommended	yes .	Expansion – Hearts of Iron IV: Man the Guns
2019-02-07	0	0	892	False	Recommended	Very good game	Expansion – Hearts

						although a bit overpriced in my opinion. I’d prefer playing the game with mods (histo…	of Iron IV: Man the Guns
2018-06-14	126	1086	676	False	Recommended	Out of all the reviews I wrote this one is probably the most serious one I wrote. For starters the c…	Dead by Daylight
2017-06-20	85	2139	612	False	Recommended	Disclaimer I Survivor main. I play games for fun not for competition so the DBD community doesn’t re…	Dead by Daylight
2016-12-12	4	55	2694	False	Recommended	ENGLISH After playing for	Dead by Daylight

						more than two years I am given the task of reviewing this game again. This …	
2017-09-17	12	228	48	False	Recommended	Out of all the reviews I wrote This one is probably the most serious one I wrote. For starters the c…	Dead by Daylight
2018-12-24	295	219	71	False	Recommended	I have never been told to kill myself more than while playing this game.	Dead by Daylight
2018-09-21	2	54	400	False	Recommended	Any longtime Dead by Daylight player knows that	Dead by Daylight

						this isn’t a horror game. If you’re looking for scar…	
2018-12-04	380	271	414	False	Recommended	if you think cs go is toxic try this game	Dead by Daylight

For training data, I will use data that I have sorted myself, and it can be seen in table provided below.

Table 4. 2 Example Training Data

Recommendation	**Review**	**Title**
Recommended	It isn't very horror and it is a very good game if you are bored! I recommend it!	Dead by Daylight
Recommended	Fun game to play with friends especially if you are a fan of the horror genre.	Dead by Daylight
Recommended	I think they did a pretty good job so far.	Dead by Daylight
Recommended	the best! I've been playing since it came out	Dead by Daylight
Recommended	Great survival horror with a fun objective. This game really gets your adrenaline flowing. The only draw back for this game is it lacks maps to play.	Dead by Daylight
Recommended	fun game	Dead by Daylight
Not Recommended	too bad optimization and game is not finished.you can die in many out physic way	PLAYERUNKNOWN'S BATTLEGROUNDS
Not Recommended	bluehole dev plans to milk this early access game before full release. dont buy it until they	PLAYERUNKNOWN'S BATTLEGROUNDS

	fix the game THEN they can start adding cosmetics.	
Recommended	I have never been told to kill myself more than while playing this game.	Dead by Daylight
Not Recommended	It's a pretty fun game too bad you can't play online with your friends.Why even include matchmaking in your game? Especially when you are not able to make it function correctly.Please do not advertise your game as co op it's inaccurate.	Dead by Daylight

4.1.2 Preprocessing Data

After all data have been imported into the *Pandas Dataframe*, the next step is data preprocessing. Preprocessing is a process for cleaning data, or better known as Data Cleaning. The purpose of this process is to make data more structured. Preprocessing has several steps, which are :

4.1.2.1 Data Lowering

The first step in preprocessing is to change all letters to lowercase. The purpose of this step is that when calculating TF-IDF, the final result does not become off because, only one uppercase can change the value of entire word.

Table 4. 3 Example of Data Lowering

Before	Great survival horror with a fun objective. This game really gets your adrenaline flowing. The only draw back for this game is it lacks maps to play.
After	great survival horror with a fun objective. this game really gets your adrenaline flowing. the only draw back for this game is it lacks maps to play.

4.1.2.2 Removing Unescaped Character

After done with lowering each data, the next step is to remove all unescaped character in each data. The purpose of this step is to remove all *html* code data. First step is to make a dictionary that contains all of unescaped character. After done with the dictionary, the next step is to remove all characters in the data that have similarities to those in dictionary.

Below is the example of unescaped data

Table 4. 4 Unescape Table

Unescape Character	
&gt;	&quot;
&lt;	&#39;
&amp;	

4.1.2.3 Stop Words Removal

Stop word is a process to remove words that lack information which have the potential to get a high value of Term Frequency.

Example:

Table 4. 5 Example of Stop Words

Example of StopWords in English	
haven’t	those
won	that
too	with
if	our
again	your

4.1.2.4 Punctuation Removal

This step aims is to remove all punctuation marks (**, ! ‘ “ : ; & #**). The goal is same as the step above, if there is a word that has a punctuation mark, it will damage values of Term Frequencies and eventually it will change the entire TF-IDF values which is not good.

Example:

Table 4. 6 Example of Punctuation Removal

Before	great survival horror with a fun objective. this game really gets your adrenaline flowing. the only draw back for this game is it lacks maps to play.
After	great survival horror with a fun objective this game really gets your adrenaline flowing the only draw back for this game is it lacks maps to play

4.1.2.5 Emoji and Emoticon Removal

This process will remove all existing emoji and emoticons in data.

4.1.2.6 URL Removal

This process will remove all existing URL in data such as, http://google.com.

4.1.2.7 Word Lemmatization

Ingason et al. [12] suggested that Lemmatization is a process to express the basic form of a word. Nirenburg [13] also said that this process aims is to normalize by returning each word to its basic form. Not only to it’s basic form, lemmatization also convert plurals into singular words.

Table 4. 7 Example of Word Lemmatization

Example of Lemmatization	
Before	great survival horror fun objective game adrenaline flowing draw back game lacks maps play
After	great survival horror fun objective game adrenaline flowing draw back game lack map play

4.1.3 Term Frequency-Inverse Document Frequency

Term Frequency Inverse Document Frequency or so-called TF-IDF is a method that used for calculating the weight of a document. This method will calculate value of Term Frequency and Inverse Document Frequency to find each data's weight.

To get the weight each data from TF-IDF will be using the formula below :

$$W = tf\ x\ idf$$

Where:

- W = Weight
- tf = term frequency each document
- idf = inverse document frequency

Inverse document frequency has its own formula. The formula for IDF shown below:

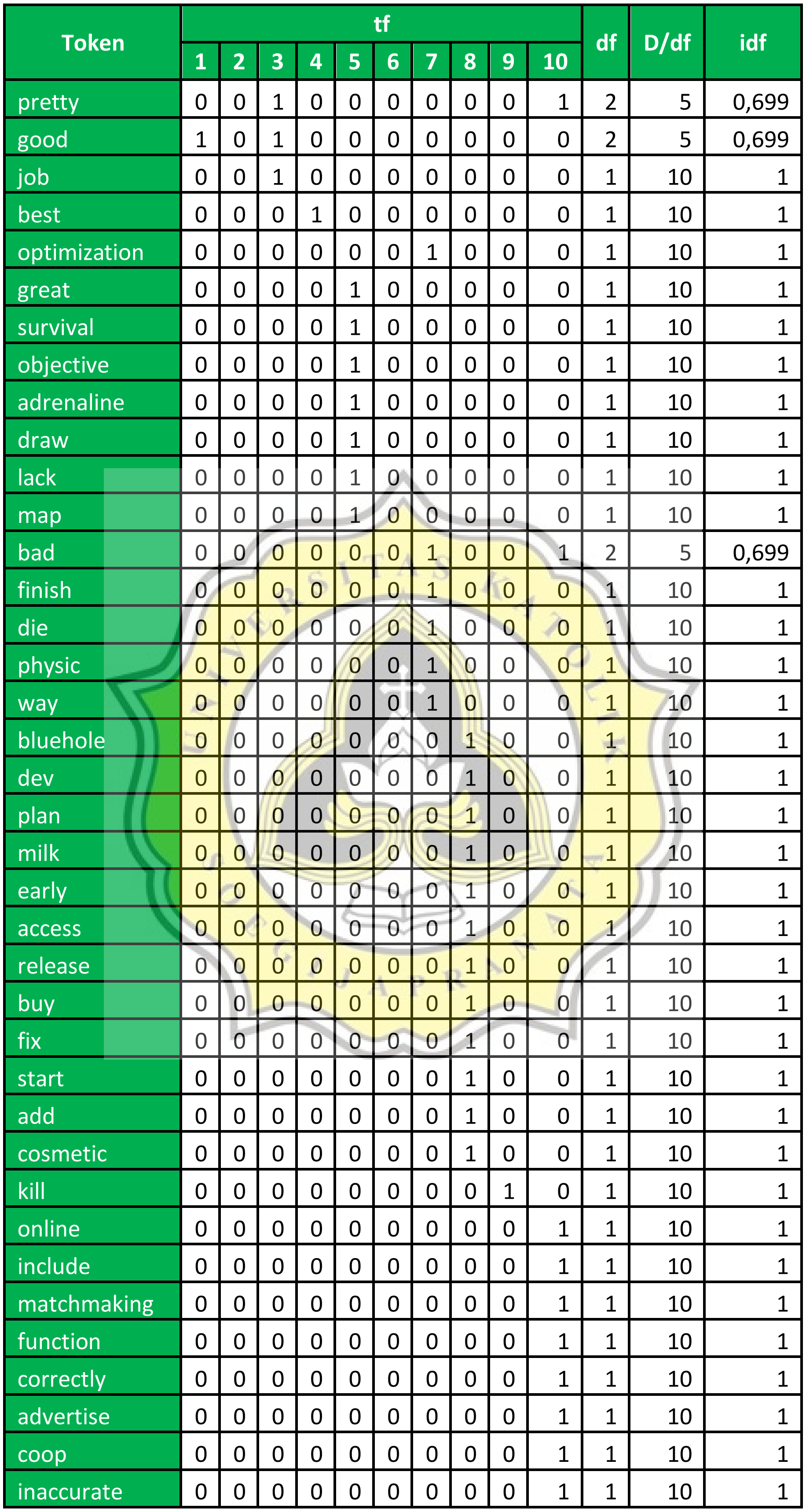
$$idf\ =\ \log\left(\frac{n}{df}\right)$$

Where:

- n = amount of the data
- df = amount of data that consist term

Table 4. 8 Table of TF and IDF

Token	tf										df	D/df	idf
	1	2	3	4	5	6	7	8	9	10			
horror	1	1	0	0	1	0	0	0	0	0	3	3,33	0,5229
good	1	0	1	0	0	0	0	0	0	0	2	5	0,699
game	1	1	0	0	2	1	1	2	1	3	8	1,25	0,0969
bored	1	0	0	0	0	0	0	0	0	0	1	10	1
recommend	1	0	0	0	0	0	0	0	0	0	1	10	1
fun	0	1	0	0	1	1	0	0	0	1	4	2,5	0,3979
play	0	1	0	1	1	0	0	0	1	1	5	2	0,301
friend	0	1	0	0	0	0	0	0	0	1	2	5	0,699
especially	0	1	0	0	0	0	0	0	0	1	2	5	0,699
fan	0	1	0	0	0	0	0	0	0	0	1	10	1
genre	0	1	0	0	0	0	0	0	0	0	1	10	1
think	0	0	1	0	0	0	0	0	0	0	1	10	1



Token	tf										df	D/df	idf
	1	2	3	4	5	6	7	8	9	10			
pretty	0	0	1	0	0	0	0	0	0	1	2	5	0,699
good	1	0	1	0	0	0	0	0	0	0	2	5	0,699
job	0	0	1	0	0	0	0	0	0	0	1	10	1
best	0	0	0	1	0	0	0	0	0	0	1	10	1
optimization	0	0	0	0	0	0	1	0	0	0	1	10	1
great	0	0	0	0	1	0	0	0	0	0	1	10	1
survival	0	0	0	0	1	0	0	0	0	0	1	10	1
objective	0	0	0	0	1	0	0	0	0	0	1	10	1
adrenaline	0	0	0	0	1	0	0	0	0	0	1	10	1
draw	0	0	0	0	1	0	0	0	0	0	1	10	1
lack	0	0	0	0	1	0	0	0	0	0	1	10	1
map	0	0	0	0	1	0	0	0	0	0	1	10	1
bad	0	0	0	0	0	0	1	0	0	1	2	5	0,699
finish	0	0	0	0	0	0	1	0	0	0	1	10	1
die	0	0	0	0	0	0	1	0	0	0	1	10	1
physic	0	0	0	0	0	0	1	0	0	0	1	10	1
way	0	0	0	0	0	0	1	0	0	0	1	10	1
bluehole	0	0	0	0	0	0	0	1	0	0	1	10	1
dev	0	0	0	0	0	0	0	1	0	0	1	10	1
plan	0	0	0	0	0	0	0	1	0	0	1	10	1
milk	0	0	0	0	0	0	0	1	0	0	1	10	1
early	0	0	0	0	0	0	0	1	0	0	1	10	1
access	0	0	0	0	0	0	0	1	0	0	1	10	1
release	0	0	0	0	0	0	0	1	0	0	1	10	1
buy	0	0	0	0	0	0	0	1	0	0	1	10	1
fix	0	0	0	0	0	0	0	1	0	0	1	10	1
start	0	0	0	0	0	0	0	1	0	0	1	10	1
add	0	0	0	0	0	0	0	1	0	0	1	10	1
cosmetic	0	0	0	0	0	0	0	1	0	0	1	10	1
kill	0	0	0	0	0	0	0	0	1	0	1	10	1
online	0	0	0	0	0	0	0	0	0	1	1	10	1
include	0	0	0	0	0	0	0	0	0	1	1	10	1
matchmaking	0	0	0	0	0	0	0	0	0	1	1	10	1
function	0	0	0	0	0	0	0	0	0	1	1	10	1
correctly	0	0	0	0	0	0	0	0	0	1	1	10	1
advertise	0	0	0	0	0	0	0	0	0	1	1	10	1
coop	0	0	0	0	0	0	0	0	0	1	1	10	1
inaccurate	0	0	0	0	0	0	0	0	0	1	1	10	1

Table 4. 9 Table of Weight each word(tf x idf)

W									
1	2	3	4	5	6	7	8	9	10
0,523	0,523	0	0	0,523	0	0	0	0	0
0,699	0	0,699	0	0	0	0	0	0	0
0,097	0,097	0	0	0,194	0,097	0,097	0,194	0,097	0,291
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0,398	0	0	0,398	0,398	0	0	0	0,398
0	0,301	0	0,301	0,301	0	0	0	0,301	0,301
0	0,699	0	0	0	0	0	0	0	0,699
0	0,699	0	0	0	0	0	0	0	0,699
0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0,699	0	0	0	0	0	0	0,699
0,699	0	0,699	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0,699	0	0	0,699
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0

W									
1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1

Table 4. 10 List of All Document’s Weights

Data	**Weight each Document**
1	4,01772877
2	4,71669877
3	4,09691001
4	1,30103000
5	8,41566878
6	0,49485002
7	5,79588002
8	12,19382003
9	1,39794001
10	11,78558006

4.1.4 Processing Data with K-Means Algorithm

The next step is classification using K-Means algorithm. In this study, the number of clusters or K will be set 2 because it was based on "Recommended" and "Not Recommended" review data. For the initial Centroid, I will use data 5 as C1 and data 10 as C2, with the following table:

Table 4. 11 Table of first Centroids

Doc	Centroid	Review	Recommendation	Weight
5	C1	great survival horror fun objective game adrenaline flowing draw game lack map play	Recommended	8,41566878
10	C2	pretty fun game bad play online friend include matchmaking game especially function correctly advertise game co op inaccurate	Not Recommended	11,78558006

The distance measurement method for K-Means used in this research is Euclidean Distance. Here's the formula for Euclidean Distance :

$$D(r,c) = \sqrt{\sum_{i=1}^{n} (r_i - c_i)^2}$$

r = review

c = centroid

With this I will start the calculation using Euclidean Distance. This process will go through several iterations, each iteration will determine which centroid has the smallest value, the centroid with the smallest value will determine the cluster. Iteration process will stop if the value of each data centroid is same as value from the previous iteration. After 1 iteration completed, the centroid value will be changed by calculating the average of total value of each cluster, in other words for Centroid 1 the average calculation will be made from total value of cluster 1, and Centroid 2 will calculate the average of total value of each data from cluster 2. The conclusion is that if the value of centroid is same as previous centroid's value, the iteration process is completed.

1st Iteration:

Value of each centroid that used in this iteration can be seen in table 4.11.

Table 4. 12 First Iteration

DATA	1	2	CLUSTER
1	4,39794001	7,76785129	1
2	3,69897000	7,06888129	1
3	4,31875876	7,68867005	1
4	7,11463878	10,48455007	1
5	0	3,36991129	1
6	7,92081875	11,29073004	1
7	2,61978876	5,98970004	1
8	3,77815125	0,40823997	2
9	7,01772877	10,38764005	1
10	3,36991129	0	2

For the next iteration centroid 1 will use the average of the sum of data 1, 2, 3, 4, 5, 6, 7, and 9. Centroid 2 will use the average summation of data 8, and 10.

Result of this calculation can be seen below.

Table 4. 13 Centroid Value from First Iteration

Centroid	Value
C1	3,779588296
C2	11,98970004

Using the Centroid values from table 4.13 I will continue for the second iteration.

2nd Iteration:

Table 4. 14 Second Iteration

DATA	1	2	CLUSTER
1	0,23814047	7,97197128	1
2	0,93711048	7,27300127	1
3	0,31732172	7,89279003	1
4	2,47855830	10,68867005	1
5	4,63608048	3,57403127	2
6	3,28473827	11,49485002	1
7	2,01629172	6,19382003	1
8	8,41423173	0,20411998	2
9	2,38164829	10,59176003	1
10	8,00599176	0,20411998	2

After this I will calculate the average of each cluster. And the results can be seen in table below.

Table 4. 15 Centroid Value from Second Iteration

Centroid	Value
C1	3,11729108
C2	10,79835629

Using this centroid value, calculation will continue for the third iteration.

3rd Iteration:

Table 4. 16 Third Iteration

DATA	1	2	CLUSTER
1	0,90043768	6,78062752	1
2	1,59940769	6,08165752	1
3	0,97961893	6,70144627	1
4	1,81626109	9,49732629	1
5	5,29837769	2,38268751	2
6	2,62244106	10,30350627	1
7	2,67858893	5,00247627	1
8	9,07652894	1,39546374	2
9	1,71935108	9,40041628	1
10	8,66828898	0,98722377	2

Table 4. 17 Centroid value from third iteration

Centroid	Value
C1	3,11729108
C2	10,79835629

From here it can be concluded that the iteration process stops at third iteration. The conclusion of this process is that the data is still not clustered properly and there are several factors that can be the cause of this, such as the use of words that are often found in other irrelevant reviews, excessive use of language expressions, and the use of sarcasm.

In this particular data, a single document can disturb the other. For example, in document 10 the reviewer giving a “Nor Recommended” review even the reviewer still wrote “pretty fun”. Because of small amount of data, one irrelevancy can disturb the result.

4.2 Design

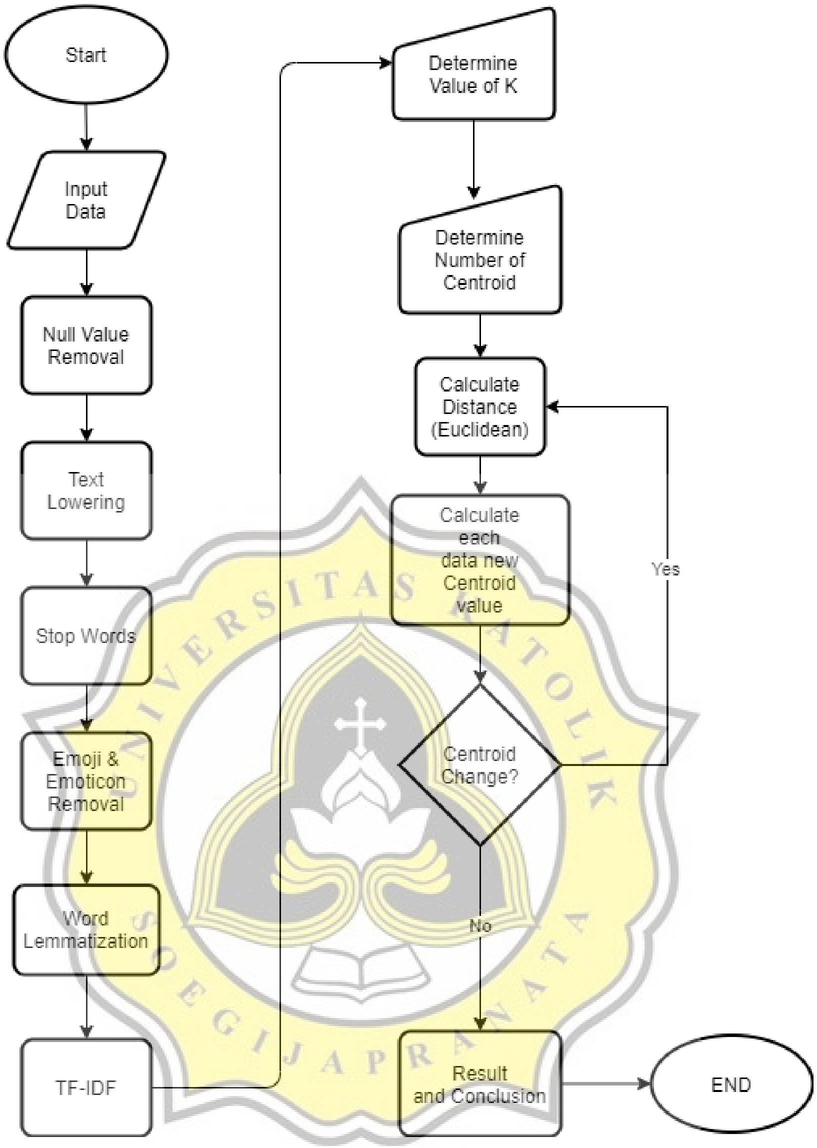


Figure 4. 1 FlowChart

The program path of this research can be summed up like the flowchart above. The first stages such as Null Value Removal, Text Lowering, Stop Words, Emoji and Emoticon Removal, and Word Lemmatization are included in the Preprocessing stage. Then proceed to the TF-IDF calculation, then after that start the calculation of the K-Means algorithm. In this process, the value of K is determined manually, then the iteration process starts and, in this process, the iteration will be repeated until the Centroid value does not change, and the clustering results are found.