CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

The main objective of this research is to observe the usage of K-Means to cluster and classify Hate Speech. Dataset is gained from www.kaggle.com/datasets.

4.1.1 Data Collecting

The data that acquired from https://www.kaggle.com contains 12 attributes, which is :

Attributes	Name
Tweet	Data from twitter which is posts from twitter
HS	Hate Speech
Abusive	Abusive but not a hate speech
HS_Individual	Hate speech to a certain individual in the context
HS_Group	Hate speech towards a certain group of organization or association
HS_Religion	Hate speech towards a certain religion or believes
HS_Race	Hate speech towards certain race or knowing
HS_Physical	Hate speech towards physical of a certain individual or a group
HS_Gender	Hate speech towards gender or sex
HS_Other	Hate speech towards attributes or other mean that related to the individual in the context
HS_Weak	Considered a weak Hate Speech
HS_Moderate	Considered a moderate Hate Speech
HS_Strong	Considered a strong Hate Speech

This project will use 1 attribute which is :

Attributes	**Name**
Tweet	Data from twitter which is posts from twitter

Below are example datas taken from the main dataset :

Table 1.1 : Data Table

No	**Data**	**HS**
1	USER USER Kaum cebong kapir udah keliatan dongoknya dari awal tambah dongok lagi hahahah'	1
2	menurutku pintu sorga ada yaitu pintu sorga yang asli dan pintu hatimu modusbanget	0
3	USER Anak pecun... ga jauh2 dr hobi zina.. haha'	1
4	Lahir dan berkembangnya gerakan komunis di Indonesia tidak dapat dipisahkan dari sebuah organisasi yg bernama ISDV'	0
5	RT USER: Bajingan Homo!!!\nHati2 terhadap anak2 kita terutama di kamar mandi laki2, jgn biarkan sendiri! URL	1
6	Apdet status bikin tugas2 kelar gak ya? Kok jancuk sekali, mbut!'	0
7	USER eh kasar kau kampang'	1
8	USER USER USER USER Gue saranin gk perlu bnyk bacot maling ayat, langsung to the point aja lu jadi orang... Yg tdk sepaham lu teriakin aja kafir sesat, liberal, pemuja dukun, bla blaaa.. pasti hatimu puas dehh ..'	1
9	Foto memang bisu ,tpi memiliki banyak arti yg sush diartikan,sma seperti kamu susah diartikan,mau putus atau lanjut ?'	0
10	Mantaaap DM dekat sama Banser bang USER	0

The HS column is to indicates whether the data is indeed Hate Speech or not. HS column is already inside the dataset.

4.1.2 Processing Data

The next step is to clean the data first . First step is Stopwords removal. This method is used to remove the unwanted words that has a function of an auxiliary verbs. Here is an example of Indonesian Stopwords :

Table 1.2 : Stopwords Table

Stopwords
Yang
Untuk
Pada
Ke
Para
Namun
Dia
Dahulu
Maka
Lagi

And the data after the stopwords removal will be as below :

Data
USER USER Kaum cebong kapir udah keliatan dongoknya awal tambah dongok hahahah'
menurutku pintu sorga yaitu pintu sorga asli pintu hatimu modusbanget
USER Anak pecun... ga jauh2 dr hobi zina.. haha'
Lahir berkembangnya gerakan komunis Indonesia dapat dipisahkan sebuah organisasi yg bernama ISDV'
RT USER: Bajingan Homo!!!\nHati2 anak2 terutama kamar mandi laki2, jgn biarkan sendiri! URL
Apdet status bikin tugas2 kelar gak ya? Kok jancuk sekali, mbut!'
USER eh kasar kau kampang'
USER USER USER USER Gue saranin gk perlu bnyk bacot maling ayat, langsung to the point aja lu jadi orang... Yg tdk sepaham lu teriakin aja kafir sesat, liberal, pemuja dukun, bla blaaa.. hatimu puas dehh ..'
Foto memang bisu ,tpi memiliki banyak arti yg sush diartikan,sma kamu susah diartikan,mau putus lanjut ?'
Mantaaap DM dekat sama Banser bang USER

But it seems that the data still have some symbols and URL that doesn't needed for the research. So it need to be cleaned as well. And here is the data after being cleaned :

Data
USER USER Kaum cebong kapir udah keliatan dongoknya awal tambah dongok hahahah
menurutku pintu sorga yaitu pintu sorga asli pintu hatimu modusbanget
USER Anak pecun... ga jauh2 dr hobi zina.. haha'
Lahir berkembangnya gerakan komunis Indonesia dapat dipisahkan sebuah organisasi yg bernama ISDV
RT USER: Bajingan Homo!!!\nHati2 anak2 terutama kamar mandi laki2, jgn biarkan sendiri!
Apdet status bikin tugas2 kelar gak ya Kok jancuk sekali mbut
USER eh kasar kau kampang
USER USER USER USER Gue saranin gk perlu bnyk bacot maling ayat langsung to the point aja lu jadi orang Yg tdk sepaham lu teriakin aja kafir sesat liberal pemuja dukun bla blaaa hatimu puas dehh
Foto memang bisu ,tpi memiliki banyak arti yg sush diartikan,sma kamu susah diartikan,mau putus lanjut ?'
Mantaaap DM dekat sama Banser bang USER

As the data shown, there are several slang words in Indonesia that needs to be formalized and simplified. As shown above, there is some “farhatabbaslaw” or “modusbanget” slang that if being tokenized will be a new word, but the formal word is “modus” and “banget”. So does the “USER” word that doesn’t needed for the research will be removed. Next step is to clean “USER” string and replace the slang. After the process, the data will be shown as below:

Data
Kaum cebong kafir sudah kelihatan dongoknya awal tambah dungu haha
menurutku pintu sorga yaitu pintu sorga asli pintu hatimu modus banget
Anak pecun tidak jauh jauh dari hobi zina haha
Lahir berkembangnya gerakan komunis Indonesia dapat dipisahkan sebuah organisasi yang bernama ISDV
Bajingan Homo nHati2 anak anak terutama kamar mandi laki laki jangan biarkan sendiri
Apdet status bikin tugas tugas selesai tidak ya Kok jancuk sekali jembut
eh kasar kamu kampang
Gue menyarankan tidak perlu banyak bacot maling ayat langsung to the point saja kamu jadi orang Yg tidak sepaham kamu teriaki saja kafir sesat liberal pemuja dukun bla blaaa hatimu puas deh
Foto memang bisu tetapi memiliki banyak arti yang susah diartikan sama kamu susah diartikan mau putus lanjut
Mantaaap DM dekat sama Banser bang

After this, the next step is to stem the data. In this research, the stemmer will be using Sastrawi Library which contains stemmer and stopwords in Indonesian Language. The result is as below :

Data
kaum cebong kafir sudah lihat dongok awal tambah dungu haha
turut pintu sorga yaitu pintu sorga asli pintu hati modus banget
anak pecun tidak jauh jauh dari hobi zina haha
lahir kembang gera komunis indonesia pisah buah organisasi nama isdv
bajing homo nhati2 anak anak utama kamar mandi laki laki jangan biar sendiri
apdet status bikin tugas tugas selesai ya kok jancuk sekali jembut
eh kasar kamu kampang
gue saran tidak perlu banyak bacot maling ayat langsung to the point saja kamu jadi orang yg tidak paham kamu riak saja kafir sesat liberal puja dukun bla blaaa hati puas deh
foto memang bisu tetapi milik banyak arti yang susah arti sama kamu susah arti mau putus lanjut
mantaaap dm dekat sama banser bang

Now that all the datas are clean, next step is to count TF-IDF. In this step, will be using help of TF-IDF library provided by sklearn library in python. After the process is done, the results is that there are 108 columns created by tokenizing word from the data. All unique words is being tokenized and counted. Here are the example from a few tokenized words :

Words
Anak
Asli
Bejat
Cebong
Maling

Before going to the step where the algorithm is used, we need to decide what data is used for the datapoint for centroid in K-Means. In this research, will be using the weight of TF-IDF in every documents for the centroid datapoint. Weight of TF-IDF is the sum of all columns that consists of TF-IDF of every words that used in a document. For example, Doc 1 until Doc 3 have the weights as per below :

Table 1.3 : TF-IDF Weight

Document	Weight
1	3.15625288
2	2.50755042
3	2.67040723

From this weight data, then can be used in K-Means algorithm.

Next step is Classification using K-Means algorithm. In this research, the K will be set 2 because the classification is consists as “Hate Speech” or “Not Hate Speech”. The C1 and C2 will be taken from data from each of classification :

Table 1.4 : Centroids

Doc	Centroid	Data	HS	Weight
1	C1	kaum cebong kafir sudah lihat dongok awal tambah dungu haha	1	3.15625288
2	C2	turut pintu sorga yaitu pintu sorga asli pintu hati modus banget	0	2.50755042

K-Means methods that will be used in this research is the Euclidian Distance. Euclidian Distance is to calculate the distance between each other datapoints based on its weight. Here is the formula for Euclidian Distance :

$$d(p,q) = \sqrt{\sum_{i=1}^{N}(p_i - q_i)^2} \quad \#(1)$$

The methods of this algorithm is going through various iteration to determine which is the cluster of which centroids. By using Euclidian Distance, the result of the distance between each centroids and datapoint will be calculated from iterations.

1st Iteration :

From the 1st iteration, the data that produced from here is shown as below :

Table 1.5 : First Iteration

No	C1	C2	Closest Cluster
1	0	0.64870246	C1
2	0.64870246	0	C2
3	0.48584565	0.16285681	C2
4	0.30480371	0.95350617	C1
5	0.02970134	0.6784038	C1
6	0.03371404	0.6824165	C1
7	1.17016393	0.52146147	C2
8	2.11038899	2.75909145	C1
9	0.18924223	0.83794469	C1
10	0.70676314	0.06207189	C2

As shown above, the closest data to each centroid is already listed. There are 7 that closest to C1 and 3 to C2. The next step is to calculate the next data of C1 and C2 based from the data above.

C1 new Value :

$$\frac{0+\ 0.30480371+\ 0.02970134+\ 0.03371404+\ 2.11038899+\ 0.18924223}{6} = 3.600894598$$

C2 new Value :

$$\frac{0 +\ 0.16285681 +\ 0.52146147 +\ 0.06207189}{4} =\ 2.402381283$$

As the results shown above, K-Means iteration will stop if the new value of centroids is the same with the previous centroid. The previous C1 is 3.15625288 and C2 is 2.50755042. And the new C1 is 3.600894598 and C2 is 2.402381283. The iteration will continue.

2nd Iteration :

The 2nd iteration data produced into the table as below :

Table 1.6 : Second Iteration

No	C1	C2	Closest Cluster
1	0.444641718	0.753871598	C1
2	1.093344178	0.105169138	C2
3	0.930487368	0.268025948	C2
4	0.139838008	1.058675308	C1
5	0.414940378	0.783572938	C1
6	0.410927678	0.787585638	C1
7	1.614805648	0.416292333	C2
8	1.665747272	2.864260588	C1
9	0.255399488	0.943113828	C1
10	1.155416068	0.043097248	C2

The new value of C1 :

$$\frac{0.444641718+ 0.139838008+ 0.414940378+ 0.410927678+ 1.665747272+ 0.255399488}{6} = 3.600894598$$

The new value of C2 :

$$\frac{0.105169138 + 0.268025948 + 0.416292333 + 0.043097248}{4} = 2.402381283$$

As the new C1 and C2 value is shown above, the new value and the previous iteration value have not changed. So the iteration stops at 2nd iteration. The result from the algorithm is that there are 3 datas that inaccurate if being crosscheck to the datasets. There are data 3, 4, 6, 7, and 9 that being clustered as a C1 which is “Hate Speech”. However the other datas are being correctly clustered to either C1 and C2 to be determined as “Hate Speech” and “Not Hate Speech”.

4.2 Design

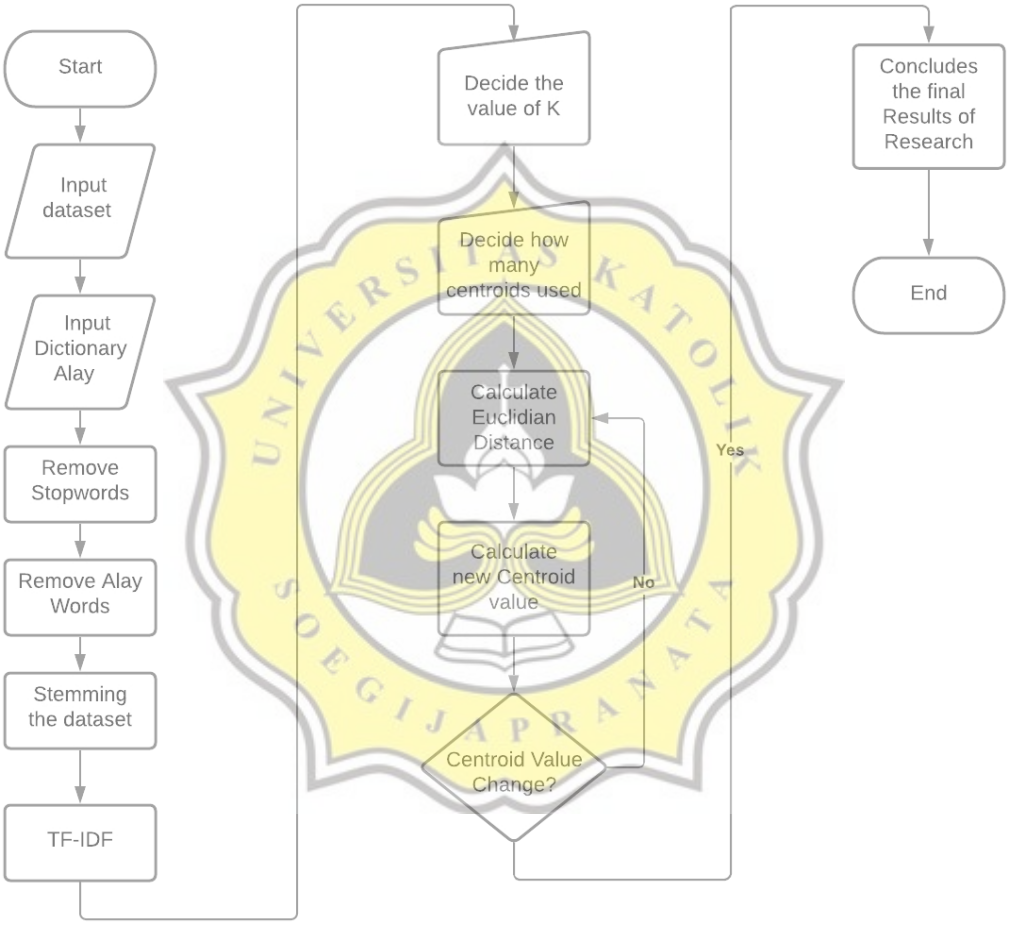

Figure 1.2. Process Flowchart

The workflow of this research can be summarized as the above flowchart. First step of the research is to input the dataset, cleaning, and stemming. Then going into the process of TF-IDF, and then the K-Means Algorithm starts. Before using the Algorithm, the K value is being decided by the classification. And then the interation process is begin. The iteration will be loop until the new value of each centroids are not changed from the previous centroids.