European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 10, 2020

Clustering Analysis from Universities in Indonesia based on Sentiment Analysis

Hendra Achmadi [1] , Isana Meranga[2], Dewi Wuisan[3], Irwan Suarly[4], I Gusti Anom Yudistira[5], Rudy Pramono[6]

[1,2,3,4,5,6] *Pelita Harapan University, Indonesia*

e-mail: [1]hendra.achmadi@uph.edu , [2]isana.meranga@uph.edu, [3]dewi.wuisan@uph.edu
[4]irwan.suarly@lecture.uph.edu, [5]anom.yudistira@gmail.com,[6]rudy.pramono@uph.edu

Abstract: There are two kind of source to determine the quality for a good university in Indonesia. First from university cluster which is publish from Ministry of Research, Technology and Higher Education issued a clustering list of Indonesian universities, the second source of data from social media, such as Twitter. In this research we use Text Mining and Data Mining Methodology to build a sentiment analysis from 50100 Tweet to assess 501 university using Python and special library in Python for Natural Language Processing a sentiment analysis , which is join the university clustering from Ministry of Research, Technology and Higher Education, so it will produce the positive, neutral and negative sentiment for each 501 universities in 2020. The next process by using R STUDIO, the process classification is continued by using K-Means, the process can be devided into two step , step 1 it will process 501 dataset university and it will build 5 cluster and secondly the similarities between Netizen cluster and cluster from Ministry of Research, Technology and Higher Education is 37 %, and step 2 after cleansing the 0 value, the result is 169 universites the similarities between Netizen cluster and cluster from Ministry of Research, Technology and Higher Education is 37 % before and after data cleansing was the same. The novelty knowledge or research finding can be derived from Netizen, firstly, the cluster can be derived based on Positive Sentiment,. Secondly, the cluster from Netizen and Cluster from Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia is only match around 37 % with cluster form Directorate General of Higher Education. And after data cleansing from 169 university was only match around 33 %..

Keywords: Sentiment Analysis, Data Mining, Clustering, K-means, Python

1. INTRODUCTION:

Higher education quality is one of the important criteria in selecting higher education institutions for high school students in Indonesia. Therefore, since 2016 the Ministry of Research, Technology and Higher Education issued a clustering list of Indonesian universities. Based on the clustering issued by the Ministry of Research, Technology and Higher Education, this could be one of the standards or benchmarks in considering the selection of higher education institutions in Indonesia.

In 2020, the Ministry of Research and Technology through the Directorate General of Higher Education issued a clustering of universities in Indonesia, while the criteria used in the assessment of higher education in Indonesia consist of four components, namely input with a weight of 20 percent, and a process with a weight of 25 percent and output with a weight. 25 percent and outcomes weighing 30 percent, and from the results of the clustering carried out

by the Directorate General of Higher Education, Ministry of Education and Culture, it was announced that there were 5 higher education clusters in Indonesia where 15 universities were included in cluster 1 and 34 universities entered into in the 2nd and 97th clusters, the tertiary institutions are included in the 3rd cluster and 400 universities are included in the 4th cluster and finally there are 1590 universities that are included in the 5th cluster, with total population is 2136 universities, based on the higher education clustering conducted by the Directorate General of Higher Education, Ministry of Education and College culture ggi has national standards, so it will help universities to evaluate the current quality and how to improve quality and for high school students have a benchmark in choosing a university based on quality, this is in line with previous research conducted by (Hendra et al. , 2019) which said that the positive influence of higher education quality on word of mouth from new students who entered college was 43.4%.

Because now high school students are in generation Z, where one of the characteristics is using social media in expressing opinions on social media, as one of the social media that is widely used in expressing opinions is Twitter, as research conducted by (Hassan Saif, Yulan He, and Harith Alani, 2012) sentiment analysis over Twitter offer organizations a fast and effective way to monitor the publics' feelings towards their brand, business, directors. And according to Hootsuite. (2020), total user Twitter in Indonesia 10,65 million people with age arrange 14-34 year which around 42,9 percent user of male and 36,1 percent famale gender in year 2020.

1.2 Research Problem

This study aims to analyses the first, what is the opinion of Twitter users on 501 university brands in Indonesia. The second is based on the sentiment analysis created from the analysis of the opinions of Twitter users, then how many college clustering will be created, then third is there a pattern of knowledge that is formed from sentiment analysis of the clustering of universities formed from Twitter users, and the fourth is what percentage of the similarity is between the clusters formed by Twitter users and the clusters formed by the Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia.

1.3 Research Question

Is there a match between netizens' perceptions in assessing the rankings of higher education, when compared with data from the Higher Education? If appropriate, you can use a marketing strategy to match the quality of clustering, otherwise netizens have their own thoughts (users tweet the senior generation, and they can have children or younger siblings, who become influencers to their families), and the university must have marketing strategy and using different approach to reach the different perception about quality from the Netizen.

2. LITERATURE REVIEW

2.1 Theory Acceptance Model (TAM)

The theory of Technology acceptance model (TAM) is formulated from the results of Ajzend an Fishben's research in 1980. TAM explains that someone when doing something will be driven by two factors, namely behavior beliefs, and normative beliefs.

These two factors encourage someone to have 2 things called outcome evaluation and motivation to comply. Outcome evaluation and motivation to comply encourage someone to behave and take personal norms. With these two things, namely Attitude and Subjective Norms, it affects attention and focus on someone in behaving which in the end will influence on a person's behavior.

Then in 1989 Davis et.al. 1989 developed TAM by examining the determinants of the use of information systems by users of information systems, and the results of his research show that the use of information systems will be influenced by interest when using information systems, then that interest will be influenced by perceptions of how useful the technology is and also by perceptions about the ease of use of that technology.

2.2 Data Mining

According to (Fatmawati, 2018) data mining is a process of finding useful new correlations, patterns or trends by mining a large number of data repositories, using pattern recognition such as statistics, and mathematical techniques, and the results of data mining can be divided into four groups. namely prediction models, clustering, association, estimation, classification.

2.3 Data Mining Processing Steps

According to (Fatmawati, 2018) said that the stages of the data mining process are starting from data selection from data sources to target data which are often referred to as datasets that are used as the basis for data processing, then the process is continued with data processing or data cleansing, here the data preparation begins for further processing, for example whether the data has number type or factor or date, and then the data in the data cleansing is also done by removing special characters, then after that the transformation is carried out, namely transforming the data from the cleansing data into the target data, the process then is to do data mining or data model based on a method that is suitable for the data, and the last is the process of interpreting the knowledge obtained from processing the data. And the data mining process stages

2.4 Text Mining and Natural Language Processing

According (Salloum et al., 2017), Text mining makes it easy to obtain a meaningful and structured data from the irregular data patterns and it used the computers to understand the unstructured data and make it structured. And the text mining is responsible for structuring the irregular data patterns written in the human language. In Text Mining using NLP (Natural Language Processing) in order to make interpretation and make categorize the kind of the unstructured message, that can be categorized to sentiment positive or negative or neutral

2.5 k-Means and Clustering

According to (Fatmawati, 2018) clustering is the process of grouping data into several clusters or groups that have maximum similarity, data between clusters has minimal similarity.

The number of clusters k must be given, and Clustering is purely descriptive. Clustering groups the items according to how similar they are. To determine the optimum k is to use the WSS method or (Withing Cluster Sum of Square), the better with the formula:

Within cluster sum of squares by cluster = between_SS / total_SS

According to (Nawrin et al. 2017), to describe the hierarchy of clusters using dendrogram. And Dendrogram is a process that captures whether the order in which clusters are merged (bottom-up view) or clusters are split (top-down view).

3. RESEARCH METHOD

The methodology used in this study is divided into two parts, the first is taking data from the opinions of Twitter users in 2020 on the names of universities in Indonesia, then using Natural Language processing using the textblob library from python programming with the sentiment analysis. function. polarity to analyze whether every tweet or opinion of twitter users on the university brand is positive or neutral or negative.

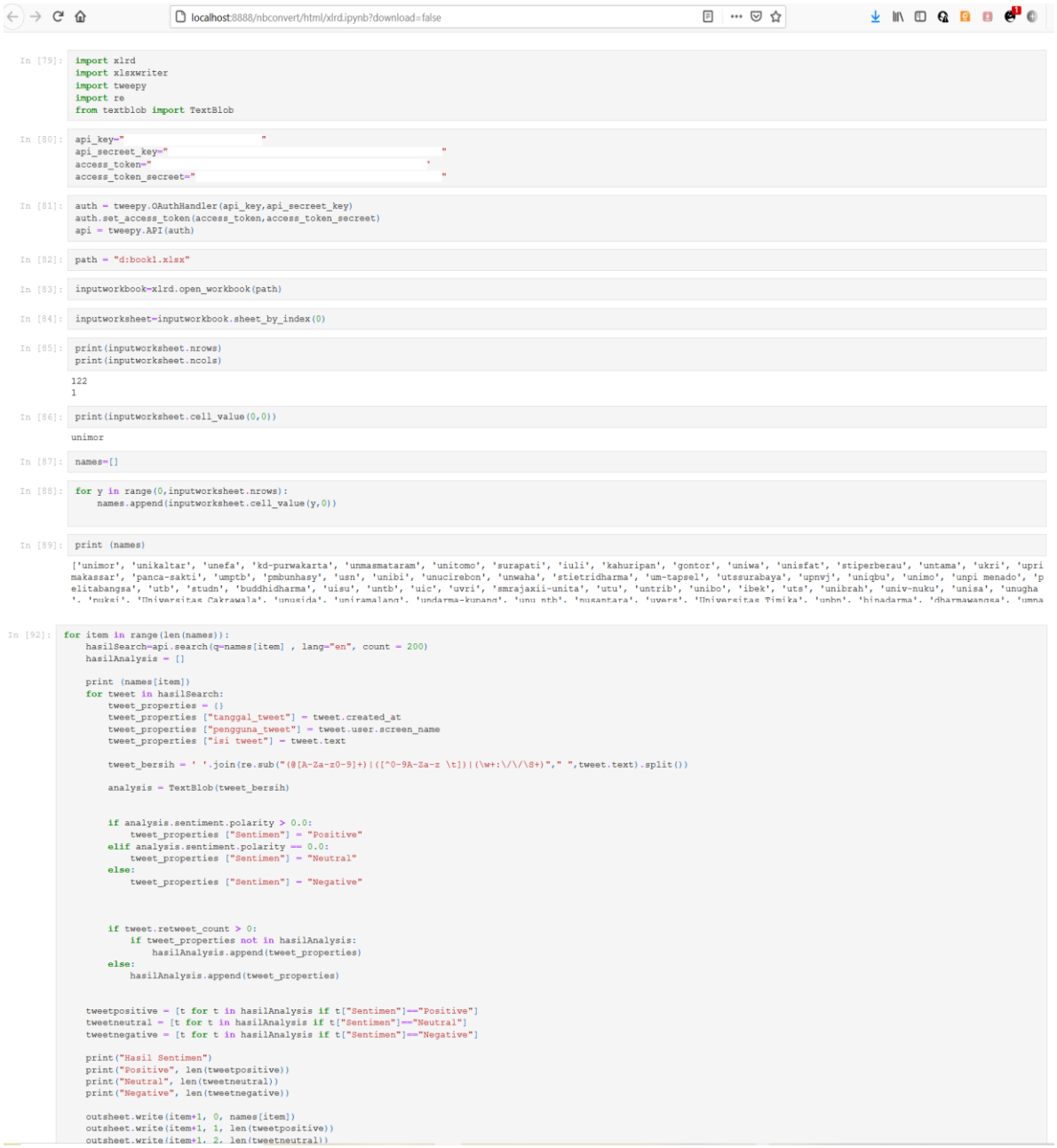
The number of tertiary institutions used to carry out the analysis is 501 universities out of a population of 2136 universities, based on the clustering made by the Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia. After getting sentiment analysis from Twitter users, the next step is to use the data mining method through 2 stages, the first is to use the existing sentiment analysis for 501 universities, then modeling by clustering 501 universities based on sentiment analysis, and then analyzed, what patterns of knowledge are formed from the sentiment analysis data, then a comparison is made between the clustering formed from the sentiment analysis of the clustering formed by the directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia, the second stage is to do cleansing of data from 501 sentiment analysis from 501 universities, which one containing 0 will be removed, then done again to test what patterns of knowledge are formed from the sentiment analysis data in stage 1 and stage 2 is the same, then a comparison is made between the clustering which is formed from sentiment analysis on the clustering formed by the Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia, what is the percentage of the similarities.

3.1. Data Collection and Methods

The main methodology that using for this research were Text Mining with Natural Language Processing by used Python and also Data Mining Methodology with clustering methodology.

Data Preparation

Data has collected from Twitter user in year 2020 by using Python programming. This program will search 100 Tweet for each university from 501 university list in Indonesia which included at the Directorate General of Higher Education, Ministry of Education and Culture of higher education clustering. So for total sample is 50100 Tweet by using Text Mining. The programming code can be seen at Figure 1. Python code for Sentiment Analysis at Twitter



```
In [79]:
import xlrd
import xlsxwriter
import tweepy
import re
from textblob import TextBlob

In [80]:
api_key="                                "
api_secreet_key="                                                "
access_token="                                                  "
access_token_secreet="                                          "

In [81]:
auth = tweepy.OAuthHandler(api_key,api_secreet_key)
auth.set_access_token(access_token,access_token_secreet)
api = tweepy.API(auth)

In [82]:
path = "d:book1.xlsx"

In [83]:
inputworkbook=xlrd.open_workbook(path)

In [84]:
inputworksheet=inputworkbook.sheet_by_index(0)

In [85]:
print(inputworksheet.nrows)
print(inputworksheet.ncols)

122
1

In [86]:
print(inputworksheet.cell_value(0,0))

unimor

In [87]:
names=[]

In [88]:
for y in range(0,inputworksheet.nrows):
    names.append(inputworksheet.cell_value(y,0))

In [89]:
print (names)

['unimor', 'unikaltar', 'unefa', 'kd-purwakarta', 'unmasmataram', 'unitomo', 'surapati', 'iuli', 'kahuripan', 'gontor', 'uniwa', 'unisfat', 'stiperberau', 'untama', 'ukri', 'upri
makassar', 'panca-sakti', 'umptb', 'pmbunhasy', 'usn', 'unibi', 'unucirebon', 'unwaha', 'stietridharma', 'um-tapsel', 'utssurabaya', 'upnvj', 'uniqbu', 'unimo', 'unpi menado', 'p
elitabangsa', 'utb', 'studn', 'buddhidharma', 'uisu', 'untb', 'uic', 'uvri', 'smrajaxii-unita', 'utu', 'untrib', 'unibo', 'ibek', 'uts', 'unibrah', 'univ-nuku', 'unisa', 'unugha
', 'puksi', 'Universitas Cakrawala', 'unusida', 'uniramalang', 'undarma-kupang', 'unu ntb', 'nusantara', 'uvars', 'Universitas Pimika', 'unbn', 'binadarma', 'dharmawangsa', 'umna

In [92]:
for item in range(len(names)):
    hasilSearch=api.search(q=names[item] , lang="en", count = 200)
    hasilAnalysis = []

    print (names[item])
    for tweet in hasilSearch:
        tweet_properties = {}
        tweet_properties ["tanggal_tweet"] = tweet.created_at
        tweet_properties ["pengguna_tweet"] = tweet.user.screen_name
        tweet_properties ["isi tweet"] = tweet.text

        tweet_bersih = ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)"," ",tweet.text).split())

        analysis = TextBlob(tweet_bersih)


        if analysis.sentiment.polarity > 0.0:
            tweet_properties ["Sentimen"] = "Positive"
        elif analysis.sentiment.polarity == 0.0:
            tweet_properties ["Sentimen"] = "Neutral"
        else:
            tweet_properties ["Sentimen"] = "Negative"


        if tweet.retweet_count > 0:
            if tweet_properties not in hasilAnalysis:
                hasilAnalysis.append(tweet_properties)
        else:
            hasilAnalysis.append(tweet_properties)


    tweetpositive = [t for t in hasilAnalysis if t["Sentimen"]=="Positive"]
    tweetneutral = [t for t in hasilAnalysis if t["Sentimen"]=="Neutral"]
    tweetnegative = [t for t in hasilAnalysis if t["Sentimen"]=="Negative"]

    print("Hasil Sentimen")
    print("Positive", len(tweetpositive))
    print("Neutral", len(tweetneutral))
    print("Negative", len(tweetnegative))

    outsheet.write(item+1, 0, names[item])
    outsheet.write(item+1, 1, len(tweetpositive))
    outsheet.write(item+1, 2, len(tweetneutral))
```

Figure 1. Python code for Sentiment Analysis at Twitter

In Figure 1, we collecting the data from Twitter, and data will be cleansing by command tweet_bersih = ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)"," ",tweet.text).split()), so text will be remove from special character, after that we preparing the data and we collected the data and saved in csv files. Sentiment Analysis from 501 universities at Indonesia .

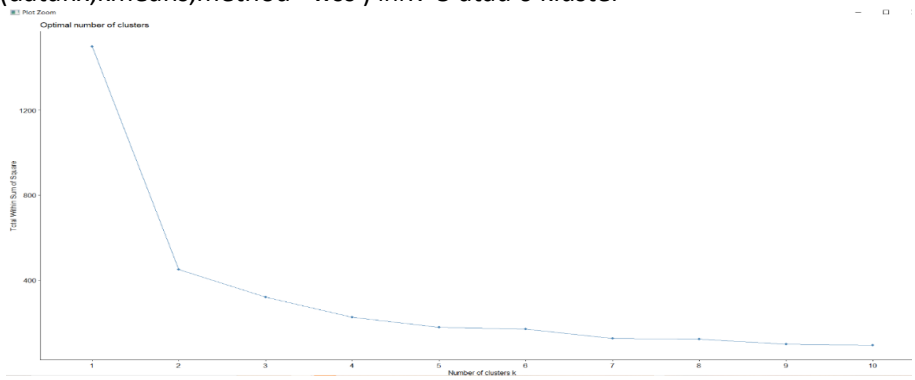
Modeling Clustering using R. Studio

Step 1 Modeling from 501 Universities

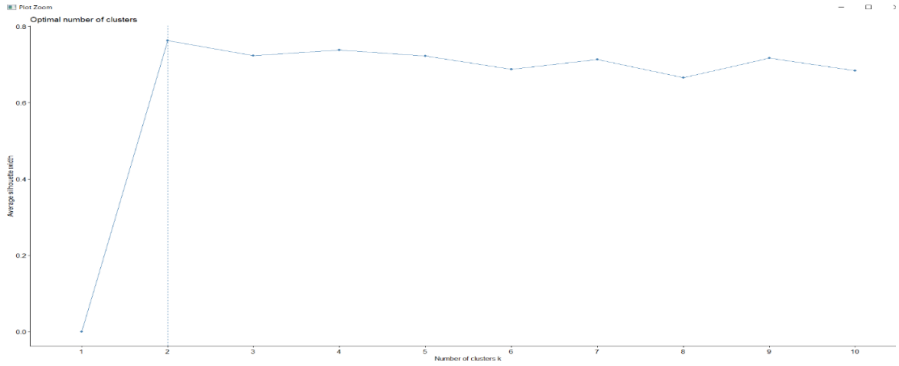
After we get the dataset , next we go to Clustering Modelling using R. Studio

```
library(ggplot2)
library(cluster)
library(factoextra)
```

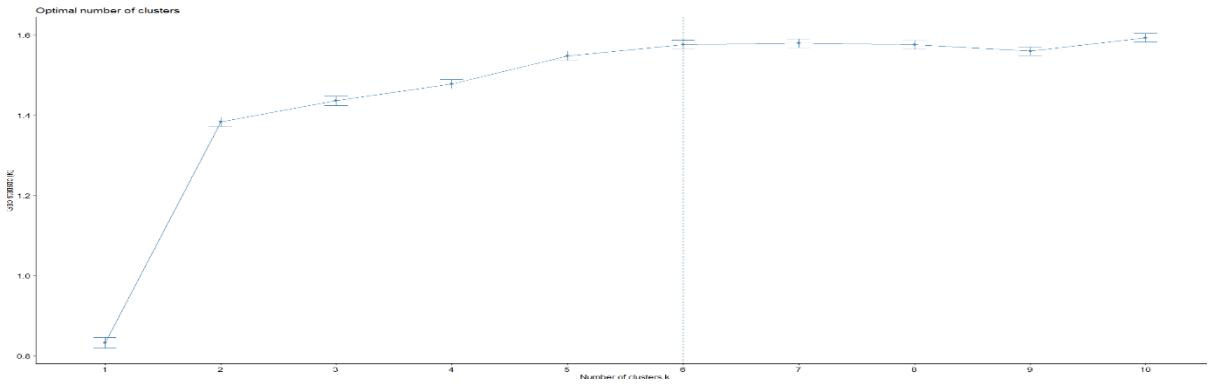
```
library(tidyverse)
univ<-read.csv2(file.choose())
univ
## clusters
numberik<-univ[2:4]
View(numberik)
dataclus<-na.omit(numberik)
datafix<-scale(dataclus)
head(datafix)
fviz_nbclust(datafix,kmeans,method='wss') ##K=5 atau 6 kluster
```




```
fviz_nbclust(datafix,kmeans,method='silhouette')## K=2
```




```
set.seed(1)
gap_stat<-clusGap(datafix,FUN=kmeans, nstart=25, K.max=10, B=501)
fviz_gap_stat(gap_stat)

Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 501)  [one "." per sample]:
.................................................. 50
.................................................. 100
.................................................. 150
.................................................. 200
.................................................. 250
.................................................. 300
.................................................. 350
.................................................. 400
.................................................. 450
.................................................. 501
. 501
```



K optimum is 6 cluster.

```
final<-kmeans(datafix,5,nstart=25)
plan<-data.frame(univ$Names,final$cluster)
plan
```

```
print(final)
K-means clustering with 5 clusters of sizes 340, 38, 67, 23, 33

Cluster means:
    Positif     Netral     Negatif
1 -0.5739056 -0.6176801 -0.49871461
2  1.2147282  1.1724490  2.73890650
3  1.0818466  1.9300535  0.89610217
4  2.8965628  0.5964684  0.19662848
5  0.2988961  0.6795699  0.02797638

Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
  3   3   4   2   2   1   1   1   2   3   5   1   3   1   5   3   1   3   2   4   2   1   1   1   5   1   1   1   3
 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  4   3   1   1   1   3   1   1   1   1   1   1   1   3   1   2   3   1   1   1   1   4   1   5   1   1   1   1   1
 59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87
  3   3   1   1   1   1   1   3   1   1   1   1   1   1   1   5   1   1   1   3   1   1   1   1   3   4   1   1   1
 88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
115 116
  1   1   1   1   1   1   1   3   1   1   1   1   1   1   1   1   1   4   1   1   4   4   1   1   1   1   1   1   3
117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
142 143 144 145
  1   2   1   1   4   1   1   2   5   1   1   1   1   1   1   1   4   1   1   1   4   1   2   5   1   3   1   3   1
146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170
171 172 173 174
  2   5   1   1   2   1   2   5   1   3   1   3   1   1   1   3   1   2   1   1   3   1   1   1   1   1   1   3   2
175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199
200 201 202 203
  1   1   1   1   2   4   5   5   1   1   3   3   2   1   4   1   1   1   3   5   1   2   5   1   5   3   1   1   1
204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228
229 230 231 232
  3   1   1   1   1   3   1   5   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   4   1   1   1
233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257
258 259 260 261
  1   1   1   1   3   3   1   1   1   1   2   1   1   1   1   1   1   1   1   1   1   3   1   1   5   2   1   1   2
262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286
287 288 289 290
  1   1   1   1   1   1   1   1   3   1   1   1   1   1   1   1   3   1   1   1   2   1   1   2   1   3   1   5   1
```

```
291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315
316 317 318 319
 1  1  5  3  1  1  1  1  1  1  3  3  5  1  1  1  1  1  1  3  1  4  3  1  4  1  1  1  1
320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344
345 346 347 348
 1  3  2  3  3  1  1  1  2  3  1  1  1  1  1  1  1  1  2  1  2  1  1  3  1  1  3  2  1
349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373
374 375 376 377
 5  5  1  1  5  3  1  1  2  1  3  1  1  1  5  1  2  3  1  2  5  1  3  1  1  1  1  4  5
378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402
403 404 405 406
 1  1  1  1  1  1  1  1  1  3  1  1  1  1  1  1  3  1  1  1  1  5  1  1  1  1  1  1  1
407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431
432 433 434 435
 1  1  1  1  2  1  1  5  1  4  1  1  2  1  3  1  3  1  1  2  1  1  1  1  1  1  1  3  1
436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460
461 462 463 464
 1  1  1  1  1  1  1  5  1  1  5  1  5  2  1  1  3  1  3  1  4  1  4  1  1  1  2  4  3
465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489
490 491 492 493
 1  4  3  2  1  1  1  1  1  4  1  1  1  1  3  2  1  5  1  3  3  2  1  1  3  1  1  5  1
494 495 496 497 498 499 501 501
 1  1  1  5  3  3  1  1
```

Within cluster sum of squares by cluster:

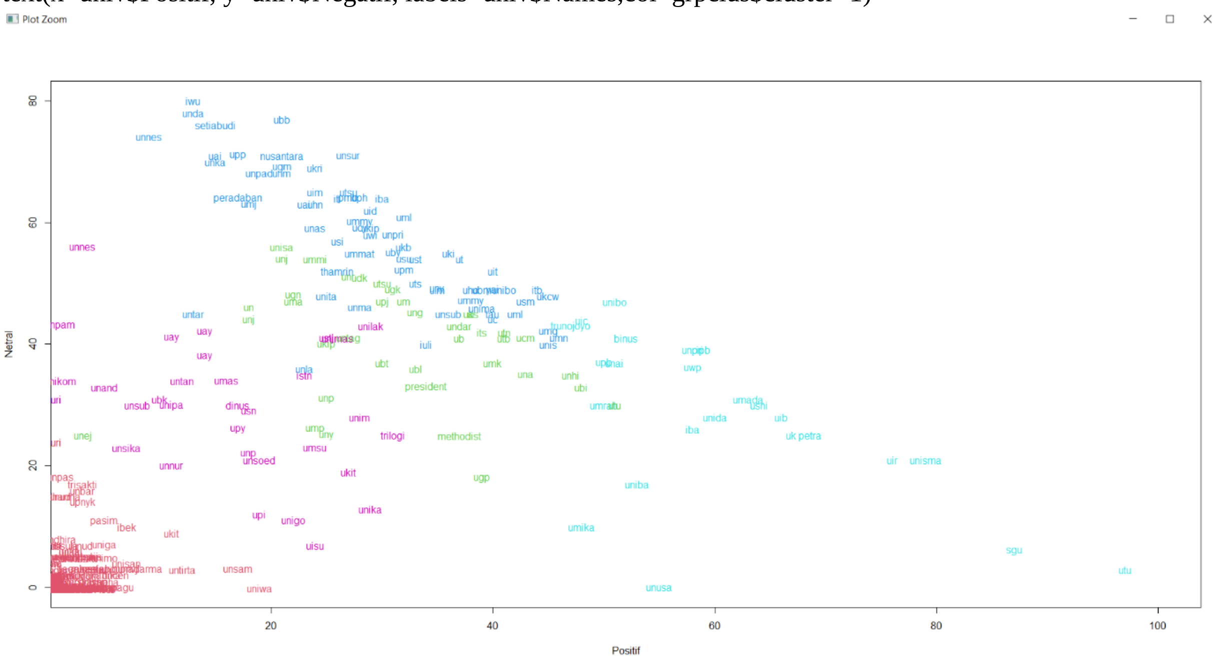
[1] 14.91832 57.03178 55.56809 28.16057 22.62714

(between_SS / total_SS = 88.1 %)

Available components:

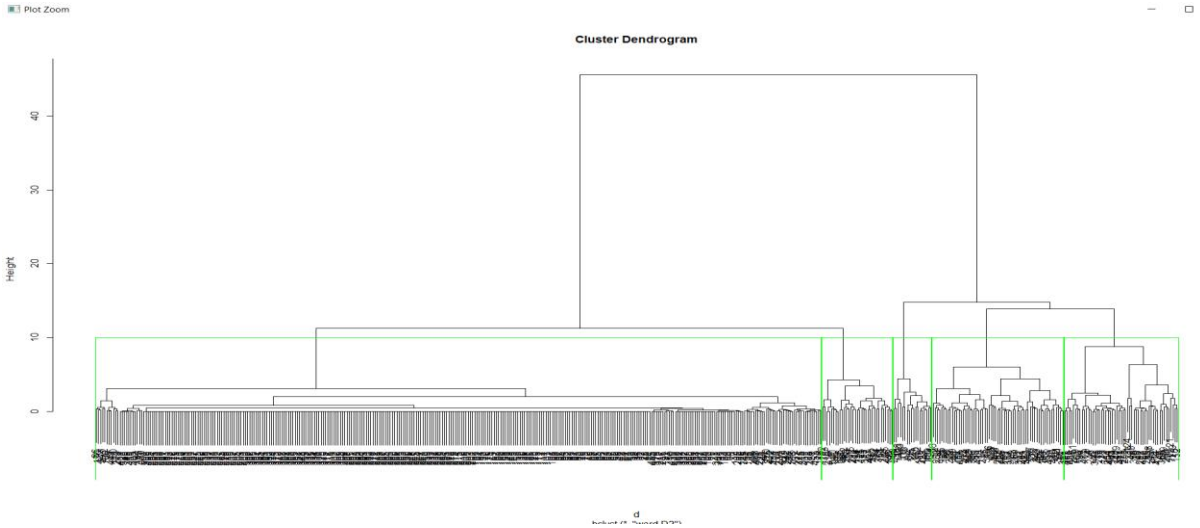
```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

```
plot(univ$Positif,univ$Netral, type="n", xlim=c(4,100), xlab="a1",ylab="a2")
text(x=univ$Positif, y=univ$Negatif, labels=univ$Names,col=grpclus$cluster+1)
```




```
univ<-read.csv2(file.choose())
```

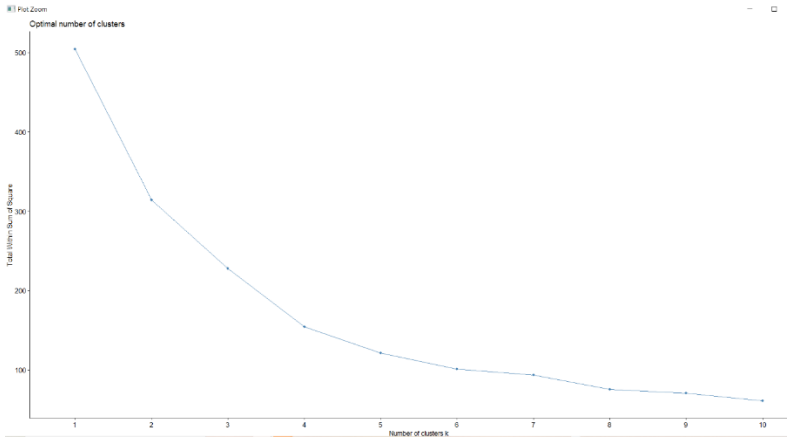
```
df<-scale(univ[-1])
d<-dist(df, method="euclidean")
hfit<-hclust(d,method="ward.D2")
plot(hfit)
grps<-cutree(hfit, k=5)
grps
rect.hclust(hfit, k=5, border="green")
```



Step 2, Modeling After Data Cleansing from Sentiment Analysis

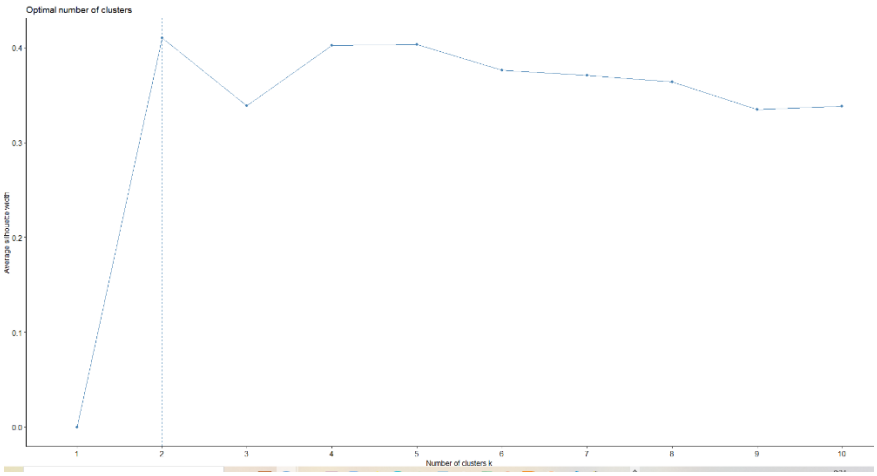
Names	Positif	Netral	Negatif
itb	44	49	7
ugm	21	69	10
ipb	59	39	2
its	39	42	19
ui	38	45	17
ub	37	41	22
unpad	19	68	13
usu	32	54	14
upi	19	12	6
uny	35	49	16
um	32	47	21
binus	52	41	7
unej	3	25	22
unp	18	22	9
unm	21	68	5
uki	36	55	9
uho	38	49	13
unas	24	59	17
ung	33	45	22
uph	28	64	8
iba	58	26	16
uay	14	38	4
ubb	21	77	2
ukip	29	59	12
uml	32	61	7
unisan	3	5	1
upnyk	3	5	2
unsub	8	30	9
utsu	27	65	8
unda	13	78	9
unpri	58	39	3
ukit	11	9	4
peradaban	17	64	14
unibo	51	47	2
unida	60	28	8
trisakti	3	17	2
untar	13	45	13
unp	25	31	24
ump	24	26	50
untan	12	34	6
uniba	53	17	5
uib	66	28	6
ubl	33	36	31
unikom	1	34	1
unla	23	36	14
uc	40	44	14
uma	22	47	31
dinus	17	30	6
unr	27	51	22
ubt	30	37	24
umsu	24	23	6
usm	43	47	10
yai	40	49	11
umn	46	41	13
unhi	47	35	18
ut	37	54	9
jayabaya	3	5	1
uniga	5	7	5
ulm	35	49	14
unj	21	54	23
uajy	1	5	1
uny	25	25	29
uk petra	68	25	7
unsoed	19	21	3
umj	18	63	2
uai	15	71	5
un	18	46	25
usni	64	30	5
ummy	28	60	12
uay	11	41	5
ubi	48	33	19
unsur	27	71	2
ummy	38	47	15
iba	30	64	6
uay	14	42	5
umada	63	31	6
unaki	2	6	2
upp	17	71	12
unis	45	40	15
undar	37	43	20
uij	4	5	1
ummat	28	55	17
ubk	10	31	3
ucm	43	41	16
ukip	25	40	35
uml	42	45	13
unbar	3	16	2
widyakart	4	2	2
unisan	7	4	1
upnyk	3	14	5
unsub	36	45	16
utsu	30	50	20
udk	28	51	21
unpri	31	58	11
ukit	27	19	3
unigo	22	11	3
upm	32	52	16
iwu	13	80	7
ucy	28	59	13
thamrin	26	52	7
umrah	50	30	4
unka	15	70	15
uby	31	55	14
ugk	31	49	20
unma	28	46	5
ukb	32	56	12
una	43	35	22
uai	23	63	7
unj	18	44	20
uks	38	45	17
jagakarsa	3	3	1
tau	40	45	15
iti	26	64	10
ugp	39	18	43
ustj	25	41	7
unika	29	13	4
unimas	26	41	1
uit	40	52	8
utn	41	42	17
ukcw	45	48	7
pasim	5	11	2
unim	28	28	8
upj	30	47	22
uim	24	65	11
ugn	22	48	30
unsika	7	23	10
uwi	29	58	13
umika	48	10	1
umas	16	34	5
iuli	34	40	6
gontor	1	2	2
ukri	24	69	3
usn	18	29	5
utb	41	41	18
studn	2	5	6
uisu	24	7	2
uic	48	44	8
utu	51	30	19
unibo	41	49	4
ibek	7	10	4
uts	33	50	17
unisa	21	56	23
nusantara	21	71	8
uika	2	6	2
upy	17	26	7
istn	23	35	10
president	34	33	17
ust	33	54	10
uhn	24	63	13
upb	50	37	13
uir	76	21	3
ummi	24	54	22
sgu	87	6	7
pmb	27	64	9
unai	51	37	12
umg	45	42	13
methodist	37	25	38
uwp	58	36	6
setiabudi	15	76	9
umk	40	37	23
trilogi	31	25	3
uid	29	62	9
unima	39	46	15
untag	27	41	32
usi	26	57	12
unnur	11	20	2
unipa	11	30	1
unita	25	48	13
ubm	39	49	12

fviz_nbclust(datafix,kmeans,method='wss') ##K=6

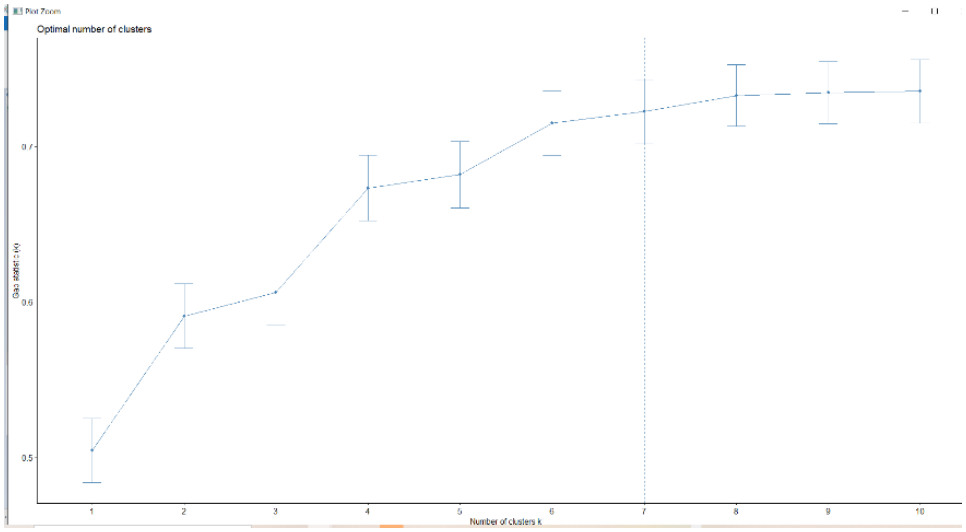


set.seed(1)

fviz_nbclust(datafix,kmeans,method='silhouette')## K=2



gap_stat<-clusGap(datafix,FUN=kmeans, nstart=25, K.max=10, B=200)



Optimal K is 7.

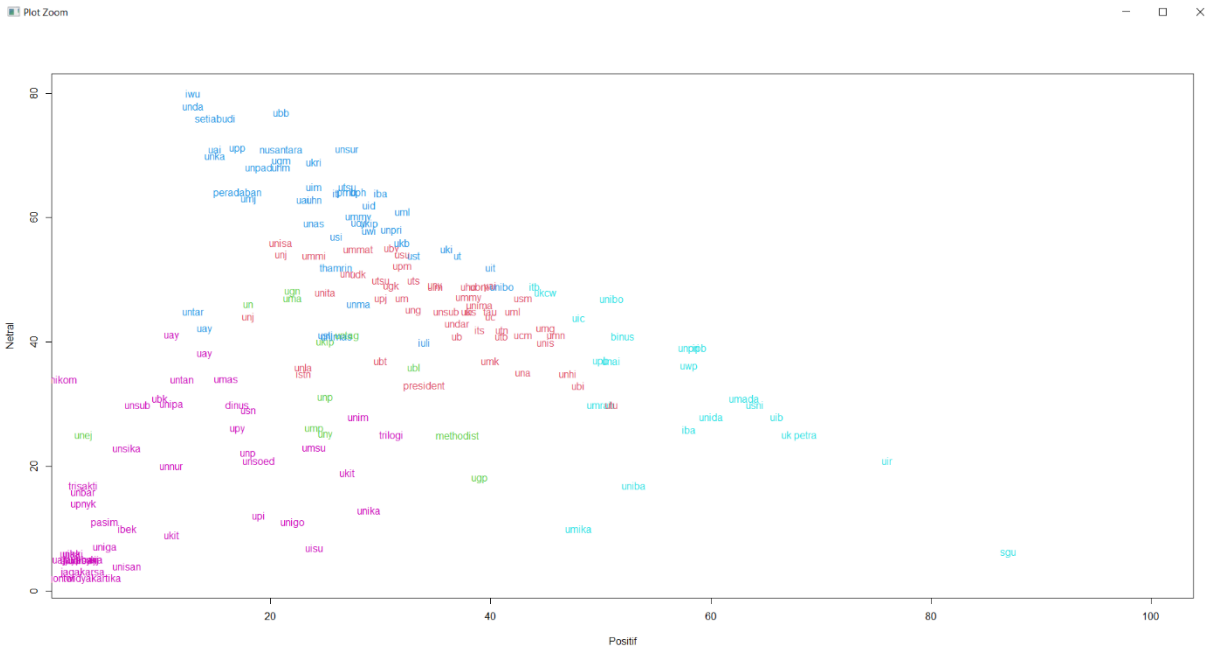
fviz_gap_stat(gap_stat)

```
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 169)  [one "." per sample]:
.................................................. 50
.................................................. 100
.................................................. 150
................... 169
```

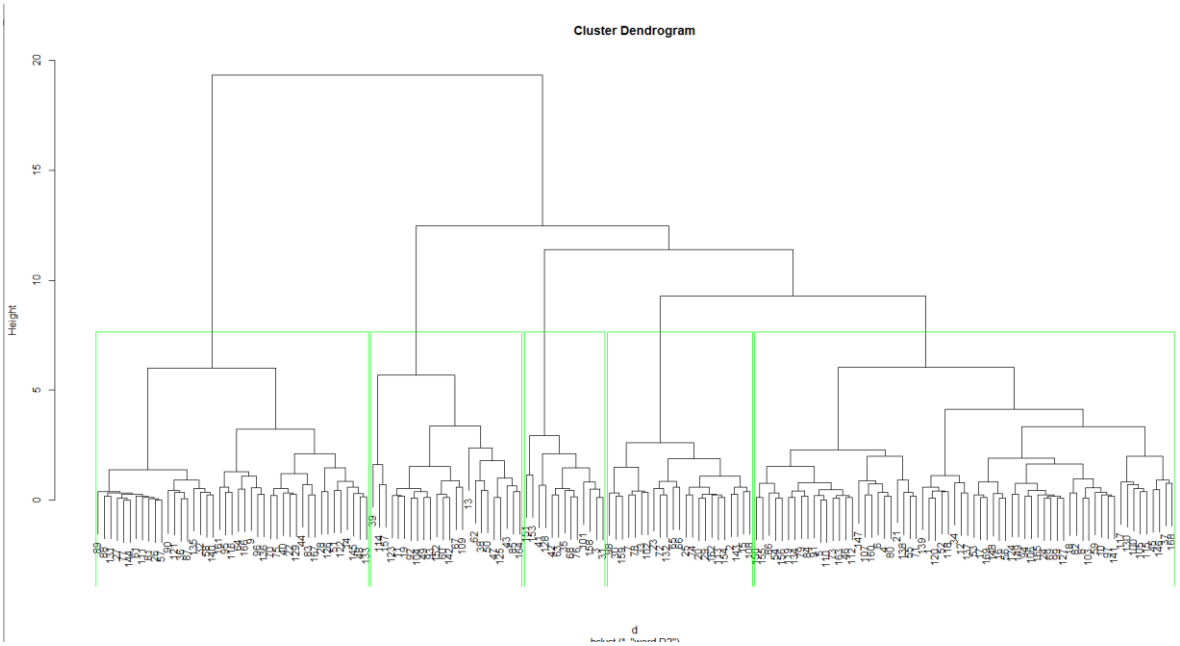
final<-kmeans(datafix,5,nstart=25)
plan<-data.frame(univ$Names,final$cluster)
plan
print(final)

	univ.Names	final.cluster
1	itb	4
2	ugm	3
3	ipb	4
4	its	1
5	ui	1
6	ub	3
7	unpad	3
8	usu	1
9	upi	5
10	uny	1
11	um	1
12	binus	4
13	unej	2
14	unp	5
15	unm	3
16	uki	3
17	uho	1
18	unas	3
19	ung	1
20	uph	3
21	iba	4
22	uay	5
23	ubb	3
24	ukip	3
25	uml	1
26	unisan	5
27	upnyk	5
28	unsub	5
29	utsu	3
30	unda	3
31	unpri	4
32	ukit	5
33	peradaban	3
34	unibo	4
35	unida	4
36	trisakti	5
37	untar	3
38	unp	2
39	ump	2
40	untan	5
41	uniba	4
42	uib	4
43	ubl	2
44	unikom	5
45	unla	1
46	uc	1
47	uma	2
48	dinus	5
49	unr	1
50	ubt	1
51	umsu	5
52	usm	1
53	yai	1
54	umn	1
55	unhi	3
56	ut	3
57	jayabaya	3
58	uniga	2
59	ulm	5
60	unj	5
61	uajy	2
62	uny	4
63	uk petra	5
64	unsoed	5
65	umj	3
66	uai	3
67	un	2
68	usni	4
69	ummy	3
70	uay	5
71	ubi	1
72	unsur	3
73	ummy	1
74	iba	3
75	uay	3
76	umada	4
77	unaki	5
78	upp	3
79	unis	1
80	undar	1
81	uij	5
82	ummat	1
83	ubk	5
84	ucm	1
85	ukip	2
86	uml	1
87	unbar	5
88	widyakartika	5
89	unisan	5
90	upnyk	5
91	unsub	1
92	utsu	1
93	udk	1
94	unpri	3
95	ukit	5
96	unigo	5
97	upm	1
98	iwu	3
99	ucy	3
100	thamrin	3
101	umrah	4
102	unka	3
103	uby	1
104	ugk	1
105	unma	3
106	ukb	3
107	una	1
108	uai	3
109	unj	1
110	uks	1
111	jagakarsa	5
112	tau	1
113	iti	3
114	ugp	2
115	ustj	3
116	unika	5
117	unimas	3
118	uit	3
119	utn	1
120	ukcw	4
121	pasim	5
122	unim	5
123	upj	1
124	uim	3
125	ugn	2
126	unsika	5
127	uwi	3
128	umika	4
129	umas	5
130	iuli	3
131	gontor	5
132	ukri	3
133	usn	5
134	utb	1
135	studn	5
136	uisu	5
137	uic	4
138	utu	1
139	unibo	3
140	ibek	5
141	uts	1
142	unisa	1
143	nusantara	3
144	uika	5
145	upy	5
146	istn	1
147	president	1
148	ust	3
149	uhn	3
150	upb	4
151	uir	4
152	ummi	1
153	sgu	4

```
154           pmb              3
155          unai              4
156           umg              1
157    methodist               2
158           uwp              4
159    setiabudi               3
160           umk              1
161      trilogi               5
162           uid              3
163         unima              1
164         untag              2
165           usi              3
166         unnur              5
167         unipa              5
168         unita              1
169           ubm              1
K-means clustering with 5 clusters of sizes 49, 12, 45, 21, 42
Cluster means:
    Positif     Netral    Negatif
1  0.4098891  0.2664229  0.6721120
2 -0.2192606 -0.3076430  2.4160365
3 -0.2165475  1.1064274 -0.3048896
4  1.7691016 -0.4205212 -0.5604012
5 -1.0680936 -1.1981261 -0.8675587

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
 4  3  4  1  1  1  3  1  5  1  1  4  2  5  3  3  1  3  1  3  4  5  3  3  3  5  5  5  3
30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
 3  4  5  3  4  4  5  3  2  2  5  4  4  2  5  1  1  2  5  1  1  5  1  1  1  1  3  5  5
59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87
 1  1  5  2  4  5  3  3  2  4  3  5  1  3  1  3  3  4  5  3  1  1  5  1  5  1  2  1  5
88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
115 116
 5  5  5  1  1  1  3  5  5  1  3  3  3  4  3  1  1  3  3  1  3  1  1  5  1  3  2  3  5
117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
142 143 144 145
 3  3  1  4  5  5  1  3  2  5  3  4  5  3  5  3  5  1  5  5  4  1  3  5  1  1  3  5  5
146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169
 1  1  3  3  4  4  1  4  3  4  1  2  4  3  1  5  3  1  2  3  5  5  1  1

Within cluster sum of squares by cluster:
[1] 25.30767 16.14312 29.33696 19.68879 30.68681
 (between_SS / total_SS =  76.0 %)

Available components:

[1] "cluster"     "centers"     "totss"       "withinss"    "tot.withinss" "betweenss"    "size"
[8] "iter"        "ifault"
>
plot(univ$Positif,univ$Netral, type="n", xlim=c(4,100), xlab="a1",ylab="a2")
text(x=univ$Positif, y=univ$Negatif, labels=univ$Names,col=grpclus$cluster+1)
```




```
univ<-read.csv2(file.choose())
df<-scale(univ[-1])
d<-dist(df, method="euclidean")
hfit<-hclust(d,method="ward.D2")
plot(hfit)
grps<-cutree(hfit, k=5)
grps
rect.hclust(hfit, k=5, border="green")
```



3.2. Results and Discussion

The result we can get from step 1, Within cluster sum of squares by cluster: [1] 14.91832 57.03178 55.56809 28.16057 22.62714, so (between_SS / total_SS = 88.1 %), and from step 2 after data cleansing, Within cluster sum of squares by cluster: [1] 25.30767 16.14312 29.33696 19.68879

30.68681 (between_SS / total_SS = 76.0 %), and then by diagram cluster dendogram, it can see the decision tree that build from clustering. Second after positif and Netral and Negatif sentiment analysis we average than it will show the new knowledge finding from table 1 Our Knowledge Finding from 501 universities before cleansing and 169 universities after cleansing, the knowledge that can be derived from Netizen, the cluster can be derived based on Positive Sentiment.

Table 1 Our Knowledge Finding from 501 Universities before cleansing and 169 Universities after Data Cleansing

501 University					After Cleansing 169 University				
Positive Sentiment	Neutral Sentiment	Negative Sentiment	Twitter user Clustering	Clustering from Directorate	Twitter user Clustering	Clustering from Directorate General of Higher	Positive Sentiment	Neutral Sentiment	Negative Sentiment
63,55	29,05	5,55	4	1	4	1	34,60	44,00	16,88
32,49	41,51	24,70	2	2	2	2	25,00	34,00	32,50
29,77	57,91	10,62	3	3	3	3	25,04	60,84	8,67
16,25	30,53	4,19	5	4	5	4	57,57	31,86	6,43
0,62	1,14	0,18	1	5	1	5	11,10	17,10	3,74
				The Mapping Cluster from					
				501 and 169 University after					
				Cleansing is The Same					

Table 2 The Percentage of Matching Perception Between Netizen Clustering and Clustering from Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia from 501 University

		Directorate General of Higher Education				
		1	2	3	4	5
	1	1	3	6	5	7
Netizen	2	5	2	8	9	14
	3	5	5	8	27	22
	4	1	2	6	11	12
	5	4	13	40	120	165

After cleansing 169 University

			Directorate General of Higher Education			
		1	2	3	4	5
	1	0	0	2	18	1
Nitizen	2	0	12	0	0	0
from	3	0	0	41	0	3
Twitter	4	19	0	0	0	22
2020	5	0	25	23	0	3

From table 2, Within cluster sum of squares by cluster = 187/501 = 37 % from 501 university before cleansing so the similarities cluster from Netizen and Cluster from Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia is only 37 %, and after cleansing from 169 university, it will get dan similarities cluster from Netizen and Cluster from Directorate General of Higher Education, Ministry of Education and Culture. From the within cluster method, if <50% is considered low.

4. CONCLUSIONS

From this research, it can be concluded that the cluster that build from the Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia is different cluster from Netizen. And after analysing it will show the new knowledge finding from table 1 Our Knowledge Finding from 501 universities, The novelty knowledge or finding research can be derived from Netizen, the cluster can be derived based on Positive Sentiment, so cluster 5 at cluster Netizen it has rank from 0,62 to 16,24 and cluster 4 at cluster Netizen is start from score 16,25 to 29,76 and cluster 3 at cluster Netizen is start from score 29,77 to 32,48, cluster 2 at cluster Netizen is start from score 32,49 to 63,54, cluster 1 at cluster Netizen is start more than score 63,54.And this pattern was the same after data cleansing.

Finally, to answer the research question, is there a match between netizens' perceptions in assessing the rankings of higher education when compared with data from the Higher Education? The cluster from Netizen and Cluster from Directorate General of Higher Education, Ministry of Education and Culture of higher education in Indonesia is only match around 37 %. And after data cleansing from 169 university was only match around 33 %. From the result the netizens had their own thoughts (users tweet the senior generation, and they can have children or younger siblings, who become influencers to their families), and the university must have marketing strategy and using different approach to reach the different perception about quality from the Netizen. Information carried out by the government is a form of government responsibility to assess the assessment of excellent primary schools, but the results of our research, the perceptions of netizens, need to be considered, and the opinions of netizens need to be considered because they represent the market, as material for determining the university's strategy for sustainability. Suggestions for further research are to find the factors that become benchmarks for netizens in assessing higher education..

5. REFERENCES

[1] Achmadi Hendra, D. P. (2019). The Influence of University Quality, Price and Service Quality on Customer Satisfaction and its Impact on the Word of Mount at the Faculty of Economics at Private Universities in Tangerang. Scholars Journal of Economics, Business and Management, 554-561.

[2] Ahmed Alsayat, H. E.-S. (2016). Social Media Analysis using Optimized K-Means Clustering. Balltimore: SERA.

[3] Hassan Saif, Y. H. (2012). Semantic Sentiment Analysis of Twitter. United Kingdom: Knowledge Media Institute, The Open University.

[4] Hootsuite. (2020). Digital 2020 Indonesia. Retrieved from https://datareportal.com: https://datareportal.com/?utm_source=Reports&utm_medium=PDF&utm_campaign=Digital_2020&utm_content=DataReportal_Promo_Slide

[5] Jiawei Han, M. K. (2012). Data Mining Concepts and Techniques . USA: Morgan Kauffman.

[6] Kiki Fatmawati, A. P. (2018). Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi. Medan: Journal of Computer Engineering System and Science.

[7] Ledolter, J. (2013). Data Mining and Business Analytics with R. Iowa: Wiley.J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[8] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.