ANALISA JEJARING SOSIAL TWITTER MENGGUNAKAN KLASTERING K-MEANS DAN HIRARKI AGGLOMERATIF

Sulastri[1], Dwi Agus Diartono[2]
Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: sul4stri66@gmail.com, dwieagus@unisbank.ac.id

ABSTRAK

Twitter merupakan salah satu media jejaring sosial yang diminati oleh masyarakat luas sampai saat ini. Biasanya twitter banyak digunakan oleh masyarakat untuk berbagi kegiatan seperti bertukar informasi ataupun tempat untuk mencurahkan isi hati. Pada twitter pengguna dapat menemukan berbagai macam informasi, bahkan ketika pengguna tidak mengetahui akun dari topik yang dicari pengguna dapat menggunakan bantuan mesin pencarian twitter untuk mencarikan informasi terkait topik yang dicari. Karena banyaknya pengguna twitter, maka akan menghasilkan kumpulan data yang besar. Hal ini ditunjukkan dari hasil survei bahwa rata-rata jumlah tweet yang ada dalam twitter adalah 600 tweet setiap detiknya pada saat kehidupan normal. Besarnya data yang dihasilkan oleh twitter, maka perlu diadakan analisa terhadap topik yang sering ditweetkan, sehingga data yang besar tersebut dapat memberikan informasi yang berguna bagi pengambil keputusan.
Penelitian ini bertujuan menganalisa data tweet yang berhubungan dengan kata kunci atau topik Piala AFF Suzuki 2016 dan Pemilihan Kepala Daerah 2017 yang sedang hangat dibicarakan di berbagai media. Analisa yang digunakan adalah analisa clustering dengan menggunakan hierarchical clustering dan k-means clustering.

***Kata kunci**: Twitter, Pilkada 2017, Paiala AFF Suzuki 2016, Clustering, K-Means, Aglomeratif*

1. PENDAHULUAN

Sebagai makhluk sosial, manusia tidak lepas dari kebutuhan dasar untuk bersosialisasi. Sosialisasi secara umum adalah proses belajar individu untuk mengenal dan menghayati norma-norma serta nilai-nilai sosial sehingga terjadi pembentukan sikap untuk berperilaku sesuai dengan tuntutan atau perilaku masyarakatnya. Salah satu cara bersosialisasi dapat dilakukan melalui komunikasi verbal maupun non verbal dan secara langsung ataupun tidak langsung. Melalui komunikasi antar individu dapat bertukar kabar atau berita yang menghasilkan suatu informasi. Di era modernisasi seperti sekarang ini, sosialisasi antar individu dapat dilakukan dengan komunikasi tidak langsung yaitu melalui media sosial. Media sosial atau sering disebut situs jejaring sosial (social network sites) adalah suatu alat (situs media online) yang dapat digunakan untuk melakukan komunikasi tanpa adanya interaksi langsung antar individu. Menurut Andreas Kaplan dan Michael Haenlein, mendefinisikan media sosial sebagai "sebuah kelompok aplikasi berbasis internet yang membangun di atas dasar ideologi dan teknologi Web 2.0 , dan yang memungkinkan penciptaan dan pertukaran "user-generated content". Terdapat banyak jenis media sosial yang berkembang sampai saat ini, salah satunya adalah situs jejaring sosial (social network sites) Twitter. Twitter merupakan situs jejaring sosial yang keberadaannya masih diminiati oleh masyarkat sampai saat ini. Twitter adalah jejaring sosial berupa blog ukuran kecil yang didirikan oleh Jack Dorsey pada bulan Maret 2006. Melalui Twitter pengguna dapat mengirim dan membaca pesan, berbagi informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan (sering disebut tweet), dengan kapasitas kata yang bisa diunggah dan ditampilkan pada timeline penggunatwitter mencapai 140 karakter. Sama halnya dengan situs jejaring sosial lain dalam Twitter disediakan suatu mesin pencarian (search engine) yang berguna untuk mempermudah pengguna dalam menemukan informasi menggunakan kata kunci. Melalui search engine pengguna dapat menemukan lebih banyak informasi yang dibutuhkan terkait topik yang ingin dicari, yaitu lebih dari satu akun yang ada di twitter.

Twitter sebagai hasil dari perkembangan teknologi informasi memungkinkan setiap waktu untuk menghasilkan kumpulan data yang banyak, dimana setiap detiknya pada saat kehidupan normal rata-rata jumlah tweet yang ada dalam twitter adalah 600 tweet. Hal tersebut tidak berlaku jika suatu waktu terjadi peristiwa-peristiwa tertentu yang menyebabkan peningkatan atau penurunan rata-rata jumlah tweet perdetiknya. Dengan adanya kumpulan data yang terus meningkat setiap waktunya yaitu berupa data tweet perlu dilakukan suatu penangan menggunakan metode khusus untuk mengalisis data pada twitter sehingga menghasilkan suatu informasi yang bermanfaat dan mengurangi kondisi yang biasa disebut "rich of data but poor of information".

Data Mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual, dimana data mining memiliki fungsi umum untuk membentuk assosiation, sequence, clustering, classification, regretion, forecasting, dan solution. Dalam data mining terdapat beberapa metode yang dapat digunakan untuk melakukan analisis data, salah satunya adalah Teks Mining.

Teks Mining didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining dimana salah satu fungsinya adalah kategorisasi.

Data mining mendukung task/fungsionalitas yang meliputi :

a. Prediktive
   Menghasilkan model berdasarkan sekumpulan data yang dapat digunakan untuk memperkirakan nilai data yang lain. Metode yang termasuk prediktive data mining :
   a. Klasifikasi : pembagian data ke dalam beberapa kelompok/kelas yang telah ditentukan sebelumnya
   b. Regresi : memetakan data ke suatu prediction variable
   c. Time Series Analysis : pengamatan perubahan nilai atribut dari waktu ke waktu
b. Deskriptive
   Mengidentifikasi pola atau hubungan dalam data untuk menghasilkan informasi baru. Metode yang termasuk deskriptive data mining :
   1) Clustering : mengelompokkan beberapa objek yang serupa ke dalam sebuah cluster, dan yang tidak serupa ke cluster yang lain
   2) Association rules : identifikasi hubungan antara data yang satu dengan yang lainnya.
   3) Summarization : pemetaan data ke dalam subset dengan deskripsi sederhana.
   4) Sequence discovery : identifikasi pola sekuensial dalam data

K-means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok/cluster. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1 , y1) , (-x-2 , y2) , …, (-x-i , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Algoritma untuk melakukan k-Means clustering adalah sebagai berikut:

a. Pilih K buah titik centroid secara acak
b. Kelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari setiap cluster merupakan titik centroid yang telah dipilih sebelumnya
c. Perbaharui nilai titik centroid
d. Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak lagi berubah

Proses pengelompokkan data ke dalam suatu cluster dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik centroid.

Hierarchical clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hierarki kelompok. Hierarchical clustering dibagi menjadi dua yaitu Agglomeratif Clustering dan Difisive Clustering. Agglomeratif Clustering mengelompokkan data dengan pendekatan bawah atas (bottom up), sedangkan Difisif Clustering menggunakan pendekatan atas bawah (top-bottom). Metode hierarchical agglomeratif clustering, mengasumsikan setiap data yang ada sebagai cluster di awal proses. Jika jumlah data adalah n, dan jumlah cluster adalah k, maka besarnya n = k. Kemudian dihitung jarak antar clusternya dengan menggunakan Euclidean distance berdasarkan jarak rata-rata antar objek. Selanjutnya, dari hasil perhitungan jarak dipilih jarak yang paling minimal dan digabungkan sehingga besarnya n = n -1. Ketika dua cluster digabungkan, jarak antara dua cluster yang digabungkan dengan cluster yang lain di-update. Penggabungan cluster akan terus dilakukan dan akan berhenti jika memenuhi kondisi jumlah k = 1. Pada akhir tahap hierarchical clustering diperoleh dendrogram yang menunjukkan urutan pengelompokan masing-masing anggota dalam cluster. Penelitian ini menggunakan metode ward sebagai metode update jarak. Metode Ward dapat membentuk cluster berdasarkan jumlah total kuadrat deviasi tiap pengamatan dari rata-rata cluster yang menjadi anggotanya [11]. Metode Ward berusaha untuk meminimalkan variasi antar objek dalam satu cluster dan memaksimalkan variasi dengan objek yang ada di cluster lainnya. Jarak antara dua cluster yang terbentuk pada metode Ward adalah sum of squares diantara dua cluster tersebut. Metode Ward didasarkan pada kriteria sum square error (SSE) dengan ukuran kehomogenan antara dua objek berdasarkan jumlah kuadrat kesalahan minimal. Perhitungan pada metode ward menggunakan rumus berikut :

$$I_{(uv)w} = \frac{n_u+n_w}{n_{uv}+n_w} I_{uw} + \frac{n_v+n_w}{n_{uv}+n_w} I_{vw} - \frac{n_w}{n_{uv}+n_w} I_{uv} \quad \ldots\ldots(2)$$

Dengan u dan v cluster yang digabung, w cluster lain yang dicari jaraknya dengan cluster gabungan uv, $I\ uv\ w$ jarak antara cluster uv dan cluster w, Iuw jarak antara cluster u dan cluster w, Ivw jarak antara cluster v dan cluster w, Iuv jarak antara cluster u dan cluster v, nu, nv, nw dan adalah banyaknya objek pada cluster ke-u, ke-v dan ke-w.

2. METODE PENELITIAN

Obyek penelitian dari penelitian ini adalah twett pada twitter tentang Pemilihan Kepala Daerah 2017 dan Piala AFF 2016.

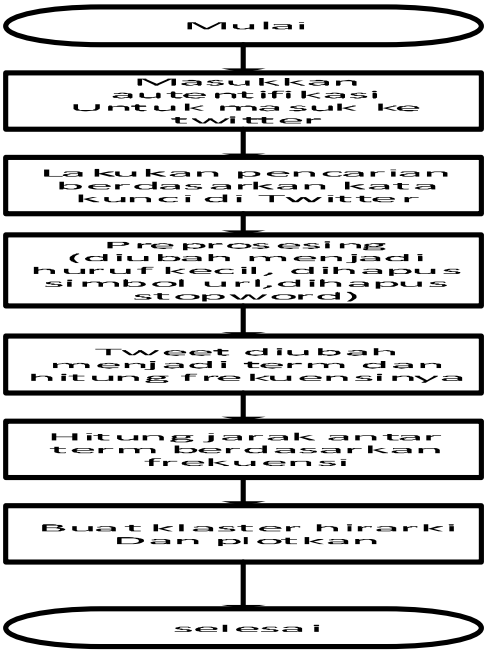
Teknik pengumpulan data yang dilakukan dalam penelitian ini adalah dengan mengambil data tweet pada twitter yang berhubungan dengan masalah Pemilihan Kepala Daerah 2017 dan Piala AFF 2016.

Metode Pengembangan Sistem :

a. Melakukan cluster terhadap twett dengan menggunakan algoritma hierarchical yang langkah-langkahnya sebagai berikut :

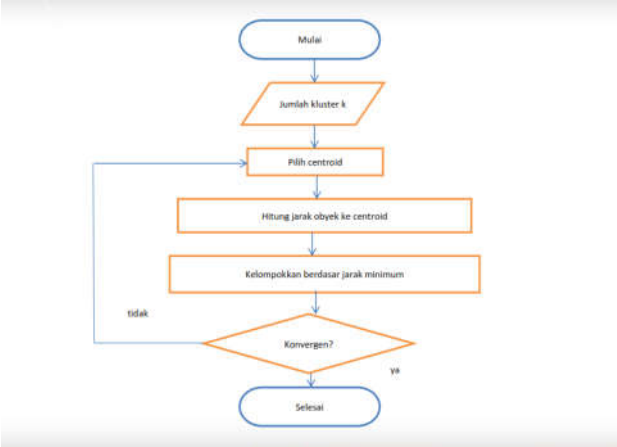
1) Tentukan k sebagai jumlah cluster yang ingin dibentuk.
2) Setiap data dianggap sebagai cluster. Kalau N= jumlah data dan n = jumlah cluater, berarti ada n=N
3) Hitung jarak antar cluster.
4) Cari 2 cluster yang mempunyai jarak antar cluste yang paling minimal dan gabungkan (berarti n=n-1)
5) Jika n > k, kembali ke langkah 3

Gambar 1 menunjukkan langkah-langkah yang dilakukan untuk mengcluster twett dengan algoritma hierarchical :



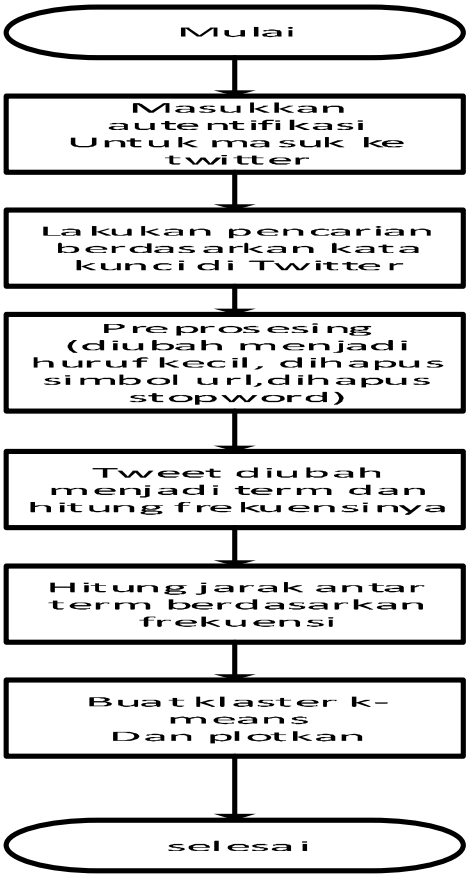
Gambar 1 Proses clustering data twett dengan Hirarchical Clustering

b. Melakukan clustering dengan menggunkan metode k-means dengan langkah-langkah sebagai berikut :



Gambar 2 Langkah proses clustering dengan k-means

Gambar 2 menunjukkan langkah clustering data twett dengan menggunkan metode k-means :



Gambar 3 Langkah mclustering data twett denga metode k-means

3. Melakukan analisa terhadap hasil clustering dengan menggunakan metode hirarchical dan metode k-means.

3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data tweet yang diperoleh dari twitter khalayak umum yang berisi mengenai Pemilihan Kepala Daerah 2017 dan Piala AFF Suzuki 2016. Cara memperolehnya adalah dengan membuat koding di bahasa R yaitu :

1. Melakukan penginstalan packages yang digunakan dalam mengambil data di twitter dan untuk keperluan pengolah serta pengklusteran yaitu :

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("tm")
install.packages("ggplot2")
install.packages("wordcloud")
install.packages("plyr")
install.packages("RTextTools")
install.packages("devtools")
install.packages("e1071")
install.packages("fpc")
install.packages("cluster")
install.packages("datasets")
```

2. Memanggil library dari packages yang diperlukan yaitu :

```
require(devtools)
library(e1071)
library(twitteR)
library(ROAuth)
library(tm)
library(ggplot2)
library(wordcloud)
library(plyr)
library(RTextTools)
```

```
library(fpc)
library(cluster)
library(datasets)
```

3 . Memanggil data twett :

```
setup_twitter_oauth("YNpg8v6N9fuTlTcCQ3VX6ONUF",
"D2pIEHseniHSKjd9O9pRZsyXbnKZxwKxNc6PVttLUAzYfUf0uU", "1712739702
KeOxsJdJUv0LQNJLsSKN95alW5iILrkRigcQaXG","3OnanvUu2ZZqWngTcGROhJJM4vq7DU23OJC
UXm1yecbdm")
tweets <- userTimeline("affsuzukicup", n = 250)
show(tweets)
n.tweet <- length(tweets)
# convert tweets to a data frame
tweets.df <- twListToDF(tweets)
```

4. Melakukan clustering dengan k-means :

```
myCorpus <- Corpus(VectorSource(tweets.df$text))
# convert to lower case
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
# remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
# remove anything other than English letters or space
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
# remove stopwords
myStopwords <- c(setdiff(stopwords('english'), c("r", "big")),"use", "see", "used", "via", "amp")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
# remove extra whitespace
myCorpus <- tm_map(myCorpus, stripWhitespace)
# keep a copy for stem completion later
myCorpusCopy <- myCorpus
term.freq <- rowSums(as.matrix(tdm))
tdm <- TermDocumentMatrix(myCorpus)
term.freq <- subset(term.freq, term.freq >= 20)
df <- data.frame(term = names(term.freq), freq = term.freq)
ggplot(df, aes(x=term, y=freq)) + geom_bar(stat="identity") +
  xlab("Terms") + ylab("Count") + coord_flip() +
  theme(axis.text=element_text(size=7))
m <- as.matrix(tdm)
# calculate the frequency of words and sort it by frequency
word.freq <- sort(rowSums(m), decreasing = T)
# colors
pal <- brewer.pal(9, "BuGn")[-(1:4)]
# plot word cloud
wordcloud(words = names(word.freq), freq = word.freq, min.freq = 3,
random.order = F, colors = pal)
#k-means clustering
d <- dist(term.freq, method="euclidian")
carsCluster <- kmeans(term.freq, 3)
clusplot(as.matrix(d), carsCluster$cluster, color=T, shade=T, labels=3, lines=0)
```

5. Melakukan clustering dengan hirarchical clustering :

```
d <- dist(t(dtms), method="euclidian")
fit <- hclust(d=d, method="ward")
fit
plot(fit, hang=-1)
library(fpc)
library(cluster)
d <- dist(t(dtm), method="euclidian")
kfit <- kmeans(d, 5)
clusplot(as.matrix(d), kfit$cluster, color=T, shade=T, labels=2, lines=0)
```

```
pamCluster <- pam(d, 3)
clusplot(as.matrix(d), pamCluster$cluster, color=T, shade=T, labels=3, lines=0)
si <- silhouette(pamCluster)
plot(si) # silhouette plot
```

Hasil Running Program

Data Twett diperoleh dari running program, datanya sebagai berikut :

Tabel 1 Hasil 250 tweet mengenai Piala AFF Suzuki 2016

No	text	Favorited	favorite Count	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource	screen Name	retweet Count	isRetweet	retweeted	longitude	latitude
1	<ed><a0><bc><ed><be><84> Merry Christmas! Here's a look back at the highlights of the #AFFSuzukiCup. Thank you fans for making thc tournam… https://t.co/EQp6OQlzzd	FALSE	88	NA	12/25/2016 3:11	TRUE	NA	8.12858E+17	NA	Twitter Web Client	affsuzukicup	105	FALSE	FALSE	NA	NA
2	<ed><a0><bc><ed><bf><86> <ed><a0><bc><ed><b7><b9><ed> <a0><bc><ed><b7><ad> Congratulations once again to five-time #AFFSuzukiCup champions #Thailand! https://t.co/CkwXJZhmoB	FALSE	67	NA	12/25/2016 3:10	FALSE	NA	8.12858E+17	NA	Twitter Web Client	affsuzukicup	82	FALSE	FALSE	NA	NA
250	Theerathon Bunmathan converts a penalty to put #Thailand 2-0 up on the night and 4-0 up aggregate! #ΛFFSuzukiCup… https://t.co/HEvJK46b4c	FALSE	32	NA	12/8/2016 13:23	TRUE	NA	8.06852E+17	NA	Twitter Web Client	affsuzukicup	64	FALSE	FALSE	NA	NA

Dari 250 tweet yang ada kemudian dilakukan proses parsing terhadap kata-kata yang muncul dan dihitung frekuensinya, hasilnya sebagai berikut :

1. Piala Aff Suzuki 2016 :

Pada masalah piala AFF Suzuki 2016 terdapat data tweet sebanyak 93 dengan 224 kata yang muncul, setelah dilakukan parsing dan dihitung frekuensi, data tersebut sebagai berikut :

Tabel 2 Hasil Parsing tweet Piala AFF Suzuki 2016

baris	kata	1	2	3	…	66	67	68	…	91	92	93
1	abduh	0	0	0	…	0	0	0	…	0	0	0
2	absen	0	0	0	…	0	0	0	…	0	1	1
3	Ada	0	0	0	…	0	0	0	…	0	1	1
4	Aff	1	1	0	…	1	1	1	…	1	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…
211	ujar	1	0	0	…	0	0	0	…	0	0	0
212	umumkan	0	0	1	…	0	0	0	…	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…
222	Yang	1	0	0	…	0	0	0	…	0	0	0
223	youtube	0	1	0	…	0	0	0	…	0	0	0
224	youtubers	0	0	0	…	0	1	1	…	0	0	0

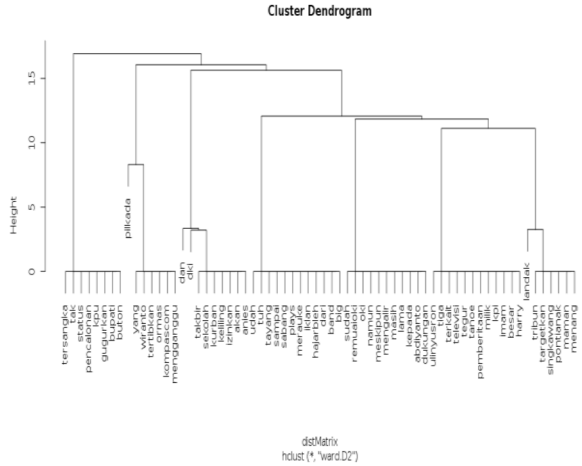
2. Pemilihan Kepala Daerah 2017 :

Pada masalah pilkada terdapat 250 twett dan dari proses ini terdapat 532 jenis kata yang muncul, setelah dilakukan parsing dan dihitung frekuensi, data tersebut sebagai berikut :

Tabel 3 Hasil Parsing data tweet PILKADA 2017

		1	2	3	……	75	76	77	……	249	250
1	Aam	0	0	0	……	0	0	0	……	0	0
2	Abdullah	0	0	0	……	0	0	0	……	0	0
3	Aceh	0	0	0	……	1	0	0	……	0	0
….	……	……	……	……	……	……	……	……	……	……	……
530	Yang	0	0	0	……	0	0	0	……	0	0
531	Zaenaldemak	0	0	0	……	0	0	0	……	0	0
532	Zuly	0	0	0	……	0	0	0	……	0	0

Hasil Clustering
Piala Aff Suzuki 2016
Hasil clustering dengan menggunakan hirarchical clustering adalah dendogram sebagai berikut :



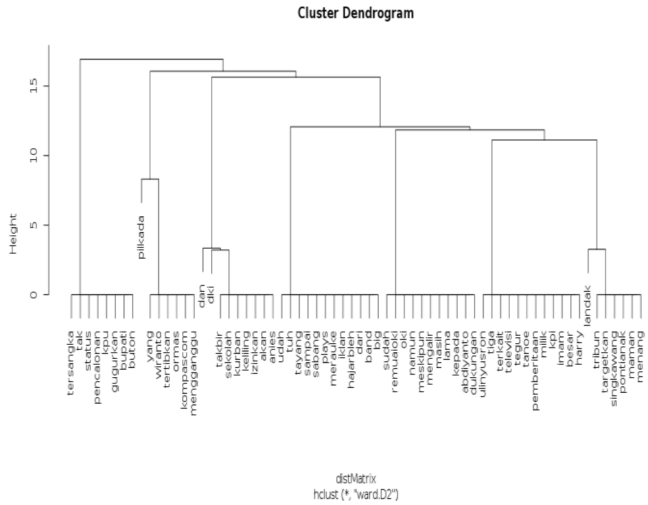
Gambar 4 Dendogram Piala AFF Suzuki 2016

Summary dari proses clustering :
K-means clustering with 3 clusters of sizes 4, 11, 8
Cluster means:
[,1]
1 59.50000
2 27.81818
3 31.62500
Clustering vector:

aff	alfred	asisten
1	3	3
bali	btw	futsal
2	2	3
indonesia	indra	ini
2	2	3
juara	panditfootball	pelatih
3	2	1
pengganti	peter	piala
2	1	1
riedl	schaller	sebagai
3	2	2
sjafri	thailand	timnas
2	2	3
umumkan	united	
2	3	

Within cluster sum of squares by cluster:
[1] 49.00000 11.63636 29.87500
(between_SS / total_SS= 97.1 %)

Pemilihan Kepala Daerah 2017 :
Hasil clustering dengan menggunakan hirarchical clustering adalah dendogram sebagai berikut :

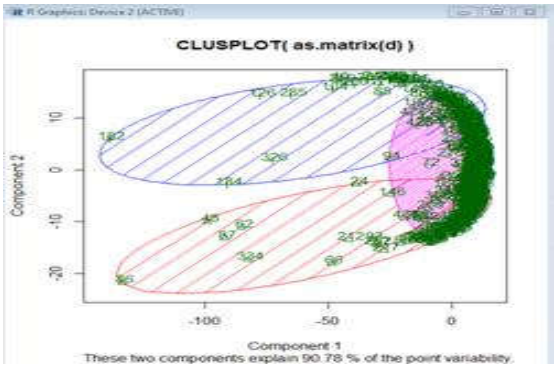


Gambar 5 Dendogram Pemilihan Kepala Daerah 2017

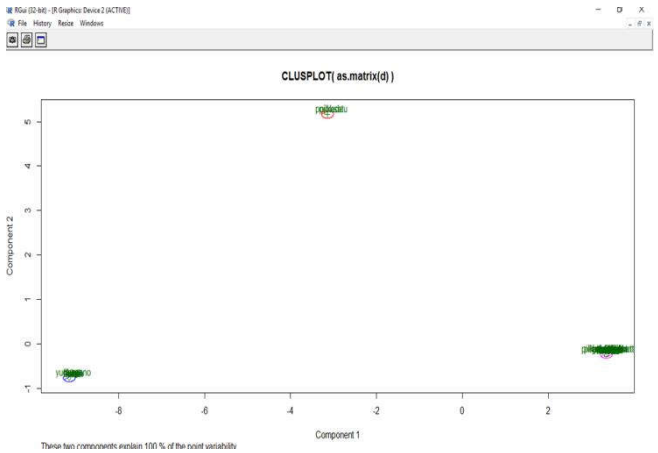
Summary dari proses clustering :
term.freq<- subset(term.freq, term.freq>= 20)
K-means clustering with 3 clusters of sizes 2, 19, 5
Cluster means: [,1]
1 218.50000
2 27.78947
3 55.20000
Clustering vector:

ahok	allah	cara	dan
2	2	3	2
dipilih	dki	gelar	haque
2	1	2	2
jakarta	jelang	koordinasi	kpu
2	2	2	3
lagi	laknatullah	marissa	melaknat
2	2	2	2
Menangkal	menyebut	pemilih	pencoblosan
3	2	3	2
perlu	pilkada	rapat	serentak
2	1	2	2
siluman	tidak		
3	2		

Within cluster sum of squares by cluster:
[1] 8064.5000 457.1579 1434.8000
(between_SS / total_SS= 86.9 %)



Gambar 6 Clustring dengan menggunakan k-means Piala AFF Suzuki 2016



Gambar 7 Clustering dengan menggunakan k-means Pemilihan Kepala Daerah 2017

Summary Clustering dengan menggunakan k-means Pemilihan Kepala Daerah 2017 :

```
K-means clustering with 3 clusters of sizes 7, 21, 2

Cluster means:
  [,1]
1    3
2    1
3    2

Clustering vector:
            agus        ahokdjarot            apakah             bukan
               1                 2                 2                 1
             cek            center             debat          februari
               2                 2                 2                 2
     halojakarta             harga             hasil       jokoanwar
               2                 1                 1                 2
             ktp              mati             nomor            online
               2                 1                 2                 2
            pada            paling           pemilih           pertama
               2                 2                 2                 2
 pilgubdkijakarta           pilkada pilkadadkijakarta          pojoksatu
               2                 3                 2                 3
          populi           sebagai             survei         terdaftar
               2                 2                 1                 2
          unggul         yudhoyono
               2                 1

Within cluster sum of squares by cluster:
[1] 0 0 0
 (between_SS / total_SS = 100.0 %)

Available components:

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

Gambar 8 Clustering dengan menggunakan k-means Pemilihan Kepala Daerah 2017

Analisa Hasil

1. Clustering untuk data tweet dengan kata kunci Piala AFF Suzuki 2016 dengan menggunakan hirarchical clustering didapat bahwa terdapat 5 kelompok besar yaitu pelatih, united, reidl, alfred dan asisten. Hal ini memberi arti bahwa pelatih sangat berperan didalam kemenangan Piala AFF Suzuki 2016, dilanjutkan dengan peran united dan pelatih tim nasional AFF sendiri yaitu Alfred Reidl serta asistennya.
2. Clustering untuk data tweet dengan kata kunci PILKADA 2017 dengan menggunakan hirarchical clustering didapat bahwa terdapat kelompok besar yaitu : pilkada, ahok, dki, kecurangan dan compascom. Hal ini memberi arti bahwa pemilihan kepala daerah sering disingkat menjadi pilkada, kemudian kata ahok banyak muncul, hal ini disebabkan karena ahok sebagai calon petahana dari pilkada di DKI Jakarta. Kata DKI juga banyak muncul karena pilkada di DKI akan segera berlangsung sehingga banyak tweet mengenai hal ini untuk menarik para kontentan. Kata berikutnya adalah kata kecurangan, hal ini muncul karena kekhawatiran publik terhadap kecurangan hasil pilkada. Kata berikutnya adalah compascom merupakan media yang paling banyak memtweet mengenai pilkada 2017.
3. Clustering untuk data tweet dengan kata kunci Piala AFF Suzuki 2016 dengan menggunakan metode k-means diperoleh bahwa data dikelopmpokan menjadi 3 cluster dengan jumlah anggota cluster 4, 11 dan 9 dengan rata-rata cluster 59.50, 27. 82 dan 31.63. Hasil ini mempunyai tingkat keyakinan 97.1 %. Hal ini menunjukkan bahwa kelompok yang terbentuk sangat meyakinkan.
4. Clustering untuk data tweet dengan kata kunci PILKADA 2017 dengan menggunakan metode k-means diperoleh 3 kelompok cluster dengan anggota 2, 19 dan 5. Rata-rat clusternya adalah 218.5, 27.79 dan 55.2 serta dengan total keyakinannya adalah 86.9%. Hal ini menunjukkan bahwa kelompok yang terbentuk cukup meyakinkan yaitu diatas 75 %.

KESIMPULAN

a. Dari hasil analisa 250 data tweet mengenai Piala AFF 2016 dengan menggunakan metode hirarchical clustering didapat bahwa terdapat 5 kelompok besar yaitu pelatih, united, reidl, alfred dan asisten. Sedangkan dengan metode k-mean clustering terdapat 3 kelompok dengan rata-rata clusternya 59.50, 27. 82 dan 31.63.
b. Dari hasil analisa 250 data tweet mengenai Pemilihan Kepala Daerah 2017 dengan menggunakan metode hirarchical clustering didapat bahwa terdapat 5 kelompok besar yaitu pilkada, ahok, dki, kecurangan dan compascom. Sedangkan dengan metode k-mean clustering terdapat 3 kelompok dengan rata-rata clusternya 59.50, 27. 82 dan 31.63.

c. Data tweet yang digunakan adalah real time jadi setiap saat clustering yang terbentuk dapat berubah, sehingga pencatuman tanggal proses pengambilan data perlu diperhatikan. Perlu dilakukan perbandingan metode mana yang paling tepat untuk melakukan clustering.

DAFTAR PUSTAKA

[1] Langgeni, D. P., Baizal, ZK. and Firdaus, A.W. 2010. *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection.* Seminar Nasional Informatika 2010 (semnasIF 2010) ISSN: 19792328. Yogyakarta

[2] Handoyo, R. 2013. *Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen*. Proposal Tugas Akhir Institut Teknologi Telkom. Bandung

[3] Arai, K., Barakbah, A. R.. 2007. *Hierarchical K-Means:an algorithm for centroids initialization for K-Means*, the Faculty of Science and Engineering, Saga University, Vol. 36, No.1

[4] Alfina, T., Santosa, B. and Barakbah, A.R. 2010. *Analisa Perbandingan Metode Hierarchical clustering, K-Means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)*. Jurnal Teknik ITS Vol. 1, (Sept, 2012) ISSN: 2301-9271. Surabaya

[5] Delen, D., Crossland, M.D. 2008. *Seeding the Survey and Analysis of Research Literature with Text mining*

[6] Turban, E. Sharda, R. Dele, D. 2011. *Decision Support and Business Intelligence Systems*. New Jersey : Pearson Education Inc.

[7] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E. 2007. *Stemming Indonesian : A ConfixStripping Approach. Transaction on Asian Langeage Information Processing*. Vol. 6, No. 4, Artikel 13. Association for Computing Machinery : New York

[8] Harlian, M. 2006. *Machine Learning Text Categorization*. University of Texas. Austin

[9] Lee, DL. 1997. *Document Ranking and the Vector-Space Model*. IEEE Software.

[10] Andayani, S. 2007. *Pembentukan Cluster dalam Knowledge Discovery in Database dengan Algoritma KMeans*. Seminar Nasional Matematika dan Pendidikan Matematika 2007. Universitas Negeri Yogyakarta. Yogyakarta.

[11] Oktavia, S., Mara, M. N., Satyahadewi, N. 2013. *Pengelompokan Kinerja Dosen Jurusan Matematika FMIPA UNTAN Berdasarkan Penilaian Mahasiswa Menggunakan Metode Ward*. Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) Volume 02, No. 2 (2013), hal 93 – 100. Tanjungpura

[12] Agusta, Y. 2007. *K-Means-Penerapan, Permasalahan dan Metode Terkait*. Jurnal Sistem dan Informatika Vol.3 , 4760.