Akuisisi dan *Clustering* Data Sosial Media Menggunakan Algoritma *K-Means* sebagai Dasar untuk Mengetahui Profil Pengguna

Binuri Ayu Dwiarni dan Budi Setiyono
Departemen Matematika, Fakultas Matematika Komputasi dan Sains Data
Institut Teknologi Sepuluh Nopember (ITS)
e-mail: masbudisetiyono@gmail.com

Abstrak*—Banyaknya informasi yang tersebar melalui sosial media menyebabkan data sosial media menjadi salah satu sumber data yang menarik untuk diteliti. Pada Twitter, hanya 11,84% dari akun protected di Twitter yang berarti sebagian besar *tweet*-nya terlihat dan mudah dibagikan. Twitter menyediakan layanan *Application Programming Interface* (*API*) yang dapat digunakan untuk mengakuisisi data. Untuk bisa menjadikan data yang diakuisi tersebut menjadi sebuah informasi yang berguna diperlukan sebuah proses salah satunya dengan *clustering*. Algoritma K-Means mempunyai kemampuan mengelompokkan data dalam jumlah cukup besar dengan waktu komputasi yang cepat dan efisien. Salah satu budaya asing yang saat ini sedang terkenal di Indonesia yaitu kebudayaan Korea Selatan (Hallyu). Oleh karena itu *Hallyu* menjadi studi kasus dalam Makalah ini. Proses Akuisisi Data Twitter dilakukan dengan cara membangun aplikasi Twitter untuk mendapatkan OAuth Twitter kemudian melakukan akuisisi dengan filter fitur-fitur yang disediakan oleh Twitter *API*. Dari hasil akuisisi data Twitter yang dilakukan dalam 5 hari dengan ujicoba 87 *keyword* "K-Pop" dan "K-Drama" dengan batasan *latitude* -4.640003 dan *longitude* 109.866141 pada radius 590km didapat 68.393 *tweet*. Hasil akuisisi data tersebut kemudian dilakukan *clustering* dengan k = 3. Dimana k_1 menunjukkan waktu *tweet* yang dianggap pagi. Sedangkan k_2 menunjukkan waktu *tweet* yang dianggap siang. Dan k_3 menunjukkan waktu *tweet* yang dianggap malam. Sehingga setelah melakukan *clustering* di dapat jam 21.00 – 01.00 merupakan mayoritas orang-orang melakukan *tweet*.

***Kata Kunci*—Akuisisi Profil Pengguna, Sosial Media, Twitter *API*, Algoritma *Clustering* K-Means, *Hallyu*.**

I. PENDAHULUAN

Di era serba digital seperti sekarang ini, teknologi dan internet sangat dibutuhkan oleh semua kalangan. Dari sebuah data statistik pengguna internet pada Januari 2018 adalah 4.021 miliar orang dengan penetrasi 53% dari total populasi di dunia yang saat itu berjumlah 7.593 milliar orang. Dari data yang sama di dapat pengguna sosial media pada Januari 2018 adalah 3.196 miliar orang. Itu berarti pengguna sosial media pada Januari 2018 adalah 42% dari populasi manusia saat itu [1]. Banyaknya informasi yang tersebar melalui sosial media menyebabkan data sosial media menjadi salah satu sumber data yang menarik untuk diteliti.

Pengertian data menurut Webster New World Dictionary, data adalah things known as assumed, yang berarti bahwa data itu sesuatu yang diketahui atau dianggap. Diketahui artinya yang sudah terjadi merupakan fakta (bukti) [2]. Namun kini, memperoleh data dapat dengan mudah dilakukan tanpa melibatkan respondennya yaitu dengan mengambil informasi data pengguna sosial media. Sosial media yang banyak digunakan sekarang ini salah satunya adalah Twitter.

Twitter adalah salah satu situs jejaring sosial online dan mikroblog yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, yang dikenal dengan sebutan kicauan (*tweet*). Pengaturan privasi data pada Twitter ada dua jenis yaitu public dan protected. Namun, hanya 11,84% akun protected di Twitter [3] yang berarti sebagian besar *tweet*-nya terlihat dan mudah dibagikan. Oleh karena itu Twitter menjadi sangat menarik untuk peneliti dalam menganalisa sebuah data. Twitter menyediakan layanan Application Programming Interface (*API*) yang dapat diintegrasikan dengan aplikasi lain. Data pada Twitter dapat diakusisi dengan memanfaatkan Twitter *API*. Untuk bisa menjadikan data yang diakuisi tersebut menjadi sebuah informasi yang berguna diperlukan sebuah proses salah satunya dengan *clustering*.

Clustering adalah metode yang mencakup pengelompokkan objek sejenis ke dalam satu cluster dan cluster yang mencakup objek kumpulan data yang dipilih untuk meminimalkan beberapa ukuran ketidaksamaan [4]. Pada data mining ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu hierarchical *clustering* dan non-hierarchical *clustering*. Salah satu contoh non-hierarchical *clustering* adalah algoritma K-Means [5]. Algoritma K-means merupakan metode *clustering* yang paling sederhana dan umum. Hal ini dikarenakan k-means mempunyai kemampuan mengelompokkan data dalam jumlah cukup besar dengan waktu komputasi yang cepat dan efisien [6].

Seiring berkembangnya zaman, pengaruh globalisasi mulai masuk ke Indonesia. Salah satu budaya asing yang saat ini sedang terkenal dan masuk ke Indonesia yaitu kebudayaan dari Korea Selatan atau yang biasa disebut dengan Hallyu atau Korean Wave. Hallyu mulai masuk ke berbagai negara dan meluas seiring berkembangnya kecanggihan teknologi dan akses internet yang mudah. Data dari The Korea Times tahun 2018 menunjukkan bahwa jumlah penggemar kebudayaan Korea di seluruh dunia meningkat 22 persen menjadi 89,19 juta, dari semula 73,12 juta penggemar pada tahun 2017.

Pada penelitian sebelumnya yang dilakukan oleh Jaka Eka Sembodo, dkk [7] yang berjudul "Data Crawling Otomatis pada Twitter". Pada penelitian tersebut membahas tentang membangun aplikasi yang berfungsi untuk crawling data Twitter secara otomatis. Akuisisi data dilakukan dengan memanfaatkan *Application Programmer Interface* (*API*) yang telah disediakan oleh Twitter. Kemudian, ada penelitian yang dilakukan oleh Dwia Pungky Arumdani [8] yang berjudul "Pengembangan Sistem Akuisisi Data Twitter Berbasis Web menggunakan Twitter Streaming *API*". Pada penelitian tersebut membahas tentang pengembangan sistem

yang dapat membantu pengguna dalam mengetahui persebaran data *tweet* serta mengetahui frekuensi hashtag yang terkandung dalam data *tweet* dan juga membandingkan hasil akuisisi Search *API* dan Streaming *API*. Selain itu ada pula penelitian yang dilakukan oleh Dwi Smaradahana Indraloka, dan Budi Santosa [9] yang berjudul "Penerapan Text Mining untuk Melakukan *Clustering* Data *Tweet* Shopee Indonesia". Pada penelitian tersebut membahas tentang akuisisi sebuah akun yaitu Shopee Indonesia yang kemudian hasil dari akuisisi tersebut dilakukan *clustering* k-means untuk mendapatkan *tweet* yang paling banyak disukai oleh pengguna Twitter yang lainnya.

Oleh karena itu, penulis membangun program aplikasi berbasis web yang dapat mengakusisi tanggal dan waktu *tweet*, lokasi, *tweet*, *retweet*, favorite, dan hashtag dengan menggunakan *keyword* yang berkaitan dengan "K-Pop dan K-Drama". Hasil akuisisi data yang didapat kemudian di cluster dengan menggunakan metode K-Means untuk mengetahui adakah keterkaitan penggemar "K-Pop dan K-Drama" dengan waktu *tweet*.

II. DASAR TEORI

A. Database

Database atau basis data kumpulan data yang disimpan secara sistematis di dalam komputer yang dapat diolah atau dimanipulasi menggunakan perangkat lunak (program aplikasi) untuk menghasilkan informasi.. Perangkat lunak yang digunakan untuk mengelola dan memanggil kueri (query) basis data disebut sistem manajemen basis data atau yang biasa dikenal dengan DBMS (Database Management System). Salah satu perangkat lunak yang sering digunakan adalah MySQL.

MySQL merupakan sistem manajemen database yang bersifat open source. MySQL merupakan sistem manajemen database yang bersifat relasional. Artinya data-data yang dikelola dalam database diletakan pada beberapa tabel yang terpisah sehingga proses manipulasi data menjadi cepat [10].

B. PHP

PHP atau Hypertext Prepocessor merupakan bahasa pemrograman yang digunakan secara luas untuk penanganan pembuatan dan pengembangan sebuah situs web dan bisa digunakan bersamaan dengan HTML. PHP adalah bahasa pemrograman server-side yang didesain untuk pengembangan web. Disebut bahasa pemrograman server-side karena PHP diproses pada komputer server. Hal ini berbeda dengan bahasa pemrograman client-side seperti JavaScript yang diproses pada web browser (client).

C. XAMPP

XAMPP adalah perangkat lunak bebas yang mendukung banyak sistem operasi yang berfungsi sebagai server yang berdiri sendiri (localhost). Nama XAMPP merupakan singkatan dari X (empat sistem operasi), Apache, MySQL, PHP dan Perl. Web server Apache berfungsi untuk simulasi pengembangan website. Melalui aplikasi ini, developer dapat menguji aplikasi secara langsung dari komputer tanpa perlu terkoneksi dengan internet. XAMPP juga dilengkapi fitur manajemen database PHPMyAdmin, sehingga pengembangan web berbasis database dapat dilakukan dengan mudah.

D. JSON dan CSV

JSON merupakan format teks yang tidak bergantung pada bahasa pemprograman apapun karena menggunakan gaya bahasa yang umum digunakan oleh programmer keluarga C termasuk C, C++, C#, Java, JavaScript, Perl, Python dll. Oleh karena sifat-sifat tersebut, menjadikan JSON ideal sebagai bahasa pertukaran-data.

Comma Separated Values atau CSV adalah suatu format data dalam basis data di mana setiap record dipisahkan dengan tanda koma (,) atau titik koma (;).

E. Geocode

Geocode merupakan proses komputasi mengubah deskripsi alamat fisik ke lokasi di permukaan bumi (representasi spasial dalam koordinat numerik). Reverse geocoding, di sisi lain, mengkonversi koordinat geografis ke deskripsi lokasi, biasanya nama tempat atau lokasi yang bisa dialamatkan. Geocoding bergantung pada representasi komputer dari titik-titik alamat, jaringan jalan / jalan, bersama dengan batas pos dan administrasi.

F. K-Pop dan K-Drama

Korean Pop (K-Pop), adalah jenis musik populer berasal dari Korea Selatan yang mempunyai boyband dan girlband. K-Pop memiliki dua unsur utama yaitu, musik dan fashion. Musik itu sendiri terbagi menjadi hiphop, rock, dan R&B yang dipadukan dengan koreografi, kostum serta wajah personil band yang menarik.

Drama Korea (K-Drama) mengacu pada drama televisi di Korea, dalam sebuah format miniseri, yang diproduksi dalam bahasa Korea. Para aktor dan aktris yang membintangi drama Korena ini mempunyai visual wajah yang menarik dengan proporsi tubuh yang sempurna dan juga akting yang sangat bagus.

G. Twitter

Twitter adalah salah satu situs jejaring sosial online dan mikroblog yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, yang dikenal dengan sebutan kicauan (*tweet*). Twitter didirikan pada bulan Maret 2006 oleh Jack Dorsey. Pada tanggal 21 Maret 2006, CEO Twitter, Jack Dorsey membuat cuitan pertamanya yang mengatakan "*just setting up my twttr*" [11]. Setelah sembilan tahun Twitter dirilis, tepatnya pada tahun 2015, rata-rata jumlah *tweet* per detik mencapai 6000 *tweet* yang berarti kira-kira 518 juta *tweet* dihasilkan dalam satu hari [12].

Terdapat fitur trending topic yang merupakan topik yang sedang banyak dibicarakan oleh pengguna Twitter secara real time. Penggunaan trending topic dengan menggunakan hashtag "#" untuk menandai *tweet* secara topikal sehingga yang lain bisa mengikuti percakapan yang berpusat pada topik tertentu. Selain itu, Twitter juga mempunyai fitur yang menampilkan apa saja *tweet* yang disukai pengguna yang disebut dengan favorite dan juga yang menampilkan apa saja *tweet* yang dibagikan kembali pengguna yang disebut dengan *retweet*.

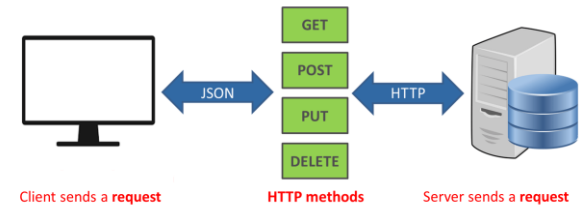
Pengaturan privasi data pada Twitter ada dua jenis yaitu public dan protected. Pengguna dapat mengikuti pengguna dengan akun public, tanpa izin yang diperlukan yang memungkinkan pengguna mengikuti dan berbagi semua *tweet* dari akun ini. Namun pada pengguna tidak dapat mengikuti dan berbagi dengan akun protected. Untuk dapat berbagi dengan akun protected, pengguna harus meminta izin dengan cara mengikutinya terlebih dahulu. Namun, hanya

11,84% dari akun protected di Twitter [3] yang berarti sebagian besar *tweet*nya terlihat dan mudah dibagikan. Sehingga Twitter menjadi sangat menarik untuk peneliti dalam menganalisa sebuah data.

Twitter memiliki layanan Application Programming Interface yang dapat diintegrasikan dengan aplikasi lain. Dengan mengunjungi https://apps.Twitter.com/ pengguna dapat memanfaatkan Twitter *API* untuk membuat aplikasi salah satunya aplikasi yang dapat mengakusisi data dengan metode autentikasi OAuth. OAuth (Open Authorization) adalah suatu protokol terbuka yang memungkinkan pengguna untuk berbagi sumber pribadi yang disimpan tanpa menyerahkan nama pengguna dan kata sandi.

H. Application Programming Interface

API adalah sekumpulan perintah, fungsi, serta protokol yang dapat digunakan oleh programmer saat membangun perangkat lunak untuk sistem operasi tertentu. *API* memungkinkan programmer untuk menggunakan fungsi standar untuk berinteraksi dengan sistem operasi [13]. *API* dilihat dari dua sisi dari developer *API* dan user (konsumer) *API* seperti pada Gambar 1. Pada sisi developer *API* yang menyediakan URL atau fungsi-fungsi apa yang dilakukan oleh *API* tersebut. Untuk dapat mengkonsumsi *API* harus menggunakan apa yang disediakan oleh developer *API* saja.



Gambar 1. Cara Kerja *API*

API tidak bergantung pada bahasa pemrograman dan tidak bergantung device apa yang digunakan [14]. *API* memungkinkan sebuah aplikasi berbicara satu sama lain tanpa sepengetahuan pengguna [15]. Misalnya pengakuisisian data pada jejaring sosial online dengan mengambil informasi pribadi yang privasinya tidak dibatasi.

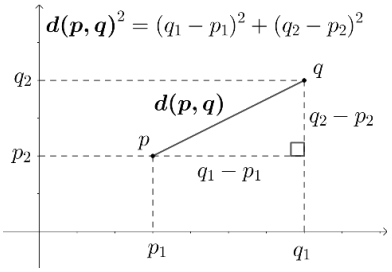
I. WEKA

Waikato Environment for Knowledge Analysis (WEKA) adalah sebuah perangkat lunak yang memiliki banyak algoritma machine learning untuk keperluan data mining. WEKA juga memiliki banyak tools untuk pengolahan data, mulai dari pre-processing, classification, regression, *clustering*, association rules, dan visualization.

J. Algoritma Clustering K-Means

Clustering adalah metode yang mencakup pengelompokkan objek sejenis ke dalam satu cluster dan cluster yang mencakup objek kumpulan data yang dipilih untuk meminimalkan beberapa ukuran ketidaksamaan[3]. Pada data mining ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu hierarchical *clustering* dan non-hierarchical *clustering* atau disebut dengan algoritma K-Means [16]. Algoritma K-means merupakan metode *clustering* yang paling sederhana dan umum. Hal ini dikarenakan k-means mempunyai kemampuan mengelompokkan data dalam jumlah cukup besar dengan waktu komputasi yang cepat dan efisien [17].

Proses *clustering* dimulai dengan mengidentifikasi data yang dikluster dengan formula Euclidean seperti pada persamaan (1) yang diilustrasikan pada Gambar 2 [19]:



Gambar 2. Ilustrasi Formula Euclidean.

$$d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \quad (1)$$

dengan:
$d(q,p)$ = jarak dari titik p ke titik q
q_i = titik q ke- i
p_i = titik pusat cluster p ke- i
i = banyaknya atribut

Suatu data menjadi anggota dari kluster ke- k apabila jarak data tersebut ke pusat kluster ke- k bernilai paling kecil jika dibandingkan dengan jarak ke pusat kluster lainnya. Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap kluster.

Nilai pusat kluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada kluster tersebut, dengan menggunakan rumus pada persamaan (2) [20].

$$\mu_k = \frac{1}{N_k}\sum_{t=1}^{Nk} x_t \quad (2)$$

dengan :
μ_k = titik centroid dari *cluster* ke- k
N_k = banyaknya data pada *cluster* ke- k
x_t = data ke- t pada cluster ke- k

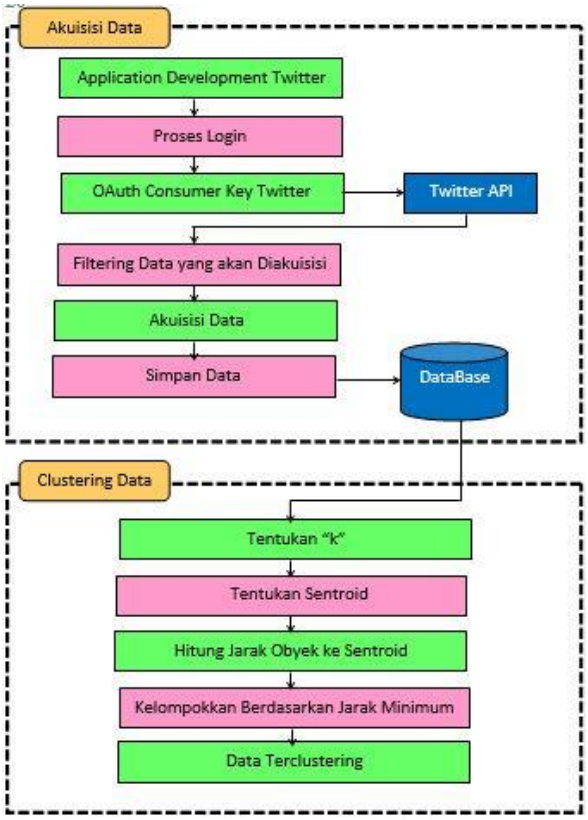
Pada umumnya K-Means *Clustering* menginisialisasi centroid secara acak[6]. Menurut Santosa [5], langkah-langkah melakukan *clustering* dengan metode K-Means *Clustering* adalah sebagai berikut:

1. Pilih jumlah klaster *k*.
2. Inisialisasi *k* pusat *cluster*. paling sering dilakukan adalah dengan cara random.
3. Alokasi semua data/ objek ke cluster terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut dan kedekatan suatu data ke klaster tertentu ditentukan jarak antara data dengan pusat klaster.
4. Menghitung kembali pusat klaster dengan keanggotaan klaster yang sekarang.
5. Menugaskan kembali setiap objek memakai pusat klaster yang baru. Jika pusat cluster tidak berubah lagi maka proses *clustering* selesai.

III. METODOLOGI PENELITIAN

Pada bab ini dibahas mengenai metode yang digunakan dalam membangun program aplikasi berbasis web yang dapat mengakusisi data twitter menggunakan algoritma K-Means.

Tahap-tahap penelitian disajikan pada Gambar 3**:**



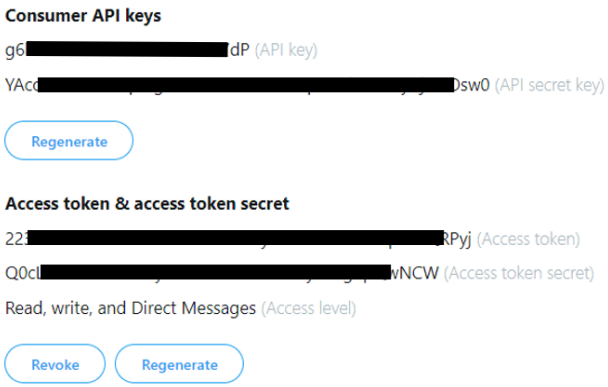
Gambar 3. Diagram Alir Tahap Analisis Data dan Pengembangan Perangkat Lunak.

IV. HASIL DAN PEMBAHASAN

A. *Uji Coba Akuisisi Data Twitter*

Pada uji coba ini, terdapat beberapa batasan. Adapun batasan dalam melakukan uji coba pada aplikasi pencarian adalah sebagai berikut:

1. Waktu pengambilan data *tweet* dilakukan selama 5 Hari yaitu pada Tanggal 20, 21, 22, 25, dan 26 Maret 2019 dengan memanfaatkan Task Scheduling pada Windows untuk melakukan crawling automatis dengan jeda 15menit.
2. Data yang dapat diproses merupakan data *tweet* dimana akun pengguna Twitter tidak diprivasi. Hasil yang didapatkan dari ujicoba ini yaitu 68.393 *Tweet*.



Gambar 4. *OAuth* pada Aplikasi BinAcquisition.

Twitter mempunyai data *user* dan data *tweet* untuk setiap penggunanya. Setiap data tersebut dapat terakuisisi dengan memanfaatkan Twitter *API*. Untuk dapat mengakusisi data Twitter maka mendaftar sebuah aplikasi Twitter. Sebelum mendaftar sebuah aplikasi Twitter, harus dilakukan login. Kemudian Aplikasi Twitter akan menyediakan OAuth *API* sebagai bentuk autentikasi. Setelah mendapatkan *OAuth API* (Consumer Key, Consumer Secret, Access Token, dan Access Secret Token) proses akuisisi data user dan data *tweet* bisa dilakukan. Contoh OAuth Twitter *API* bisa dilihat pada Gambar 4.

Setelah mendapatkan OAuth *API*, dilakukan crawling data untuk mengakuisisi data berdasarkan lokasi Geocode, dan input *keyword*. Pada Makalah ini *keyword* yang diinputkan berkaitan dengan "K-Pop" dan "K-Drama" yang berjumlah 87 *keyword* yang dapat dilihat pada Tabel 1.

Tabel 1.
Keyword yang diinputkan

2PM	Ha Ji Won	JYP Entertainment	Lee Young Ae	Shin Ye Eun
Ahn Jae Hyun	Han Ga In	Kang Ha Neul	Mamamoo	Shinee
Apink	Han Hyo Joo	Kim GO Eun	Momoland	SM Entertainment
Astro	Hyun Bin	Kim Ji Won	Nam Joo Hyuk	SNSD
Bae Suzy	iKON	Kim Kwon	NCT	So Ji Sub
Bigbang	Im Yoon Ah	Kim So Hyun	Nuest	Son Ye Jin
Black Pink	Itzy	Kim Soo Hyun	Park Bo Gum	Song Hye Kyo
BTOB	iZone	Kim Tae Hee	Park Bo Young	Song Ji Hyo
BTS	Jang Nara	Kim Woo Bin	Park Hae Jin	Song Joong Ki
Do Kyung Soo	Ji Chang Wook	Krystal Jung	Park Hyung Sik	Super Junior
EXO	Ji Sung	Lee Dong Wook	Park Min Young	TVXQ
G-Dragon	Jin Young	Lee Jong Suk	Park Seo Joon	Twice
Girl Generation	Jo In Sung	Lee Joon Gi	RedVelvet	WannaOne
Go Ara	Jo Jung Suk	Lee Min Ho	Seo In Suk	Won Bin
Go Hye Sun	Jun Ji Hyun	Lee Min Ki	Seo Kang Joon	YG Entertainment
Gong Hyo Jin	Jun Hae In	Lee Seung Gi	Seungri	Yoo Seung Ho
Gong Yoo	Jung Il Woo	Lee Seung Kyung	Shin Min A	Yoon Eun Hye
Got7	Jung Joon Young			

Sedangkan untuk lokasi geocode dibatasi dengan *latitude* (-4.640003) dan *longitude* (109.866141) dengan radius 590KM seperti pada Gambar 5. Batasan lokasi *tweet* ini dilakukan supaya waktu *tweet* yang terdapat pada *tweet* diambil menjadi *realtime* dimana hanya pada Waktu Indonesia Barat.

Gambar 5. Peta Batasan Lokasi Geocode.

Batas pencarian untuk mendapatkan *Tweet* dalam jumlah besar / GET status hingga 100 per panggilan. Untuk fitur data yang akan digunakan untuk Makalah ini adalah ID User, Akun User, Lokasi User, dan Waktu *Tweet*.

Pada proses *crawling tweet*, Twitter melakukan payload yang artinya hasil yang dikirim kembali setelah *API* panggilan. Pada Twitter menggunakan *payload* berupa *JSON*. Ini juga dapat disebut ini *output*, atau set hasil. *Payload* ini biasanya terbuat dari objek, yang mewakili konsep-konsep di Twitter seperti pengguna, status, dan lainnya. Pada Gambar 6 menunjukan potongan JSON fitur data *tweet* dengan *keyword* "Song Joong Ki". Sedangkan pada Gambar 7 menunjukan potongan JSON fitur data user dengan *keyword* "Song Joong Ki".


```
"statuses":[
      {"created_at":"Sun Jun 16
04:30:44 +0000 2019",
      "id":"1140114463053365248",
      "text":"@iwonderwoo @svthingy
Tapi aku suka sih sama idola yg gaptek
gitu kek kai, song joong ki, Mark yeo
jin goo",
      "truncated":false,
      "entities":{"hashtags":[],
```


Gambar 6. Potongan JSON Fitur Data Tweet.


```
"user":{"id":"1019968434711359489",
"id_str":"1019968434711359489",
"name":"Yeol",
"screen_name":"yeoliee_park",
"location":"Bekasi Utara, Indonesia",
"description":"Hello Dear\n19 y.o",
"url":null,
"entities":{"description":{
      "urls":[]}},
      "protected":false,
      "followers_count":79,
      "friends_count":139,
      "created_at":"Thu Jul 19 15:33:19
+0000 2018",
      "favourites_count":476,
      "statuses_count":1646,
      "lang":"id",
      "profile_image_url":"http:\/\/pbs
.twimg.com\/profile_images\/11212206846
53293568\/tZ61GyEQ normal.jpg",
```

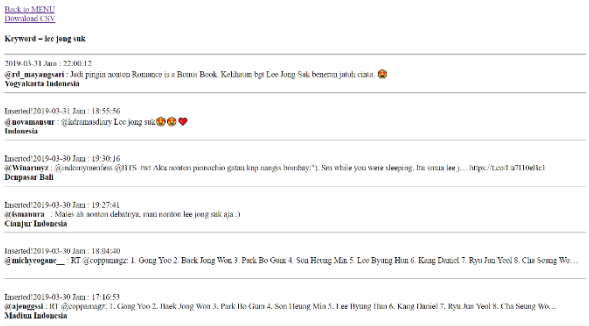

Gambar 7. Potongan JSON Fitur Data User

Karena pada Twitter format date and time menggunakan format dasar UCT atau GMT+0 maka perlu diubah menjadi waktu *realtime tweet* berdasarkan lokasi *tweet*. Pada Studi ini lokasi *tweet* merupakan lokasi Waktu Indonesia Barat (WIB) atau GMT + 7 yang diwakilkan oleh Asia/Jakarta. Pada Gambar 8 menunjukkan *tweet* dari pengguna yang ditampilkan oleh Twitter. Sedangkan pada Gambar 9 menunjukkan informasi *user* yang ditampilkan oleh Twitter. Waktu yang ditampilkan oleh Twitter merupakan sudah dalam Waktu Indonesia Barat (WIB).

Gambar 8. Tweet yang ditampilkan oleh Twitter.

Gambar 9. Info User yang ditampilkan oleh Twitter.

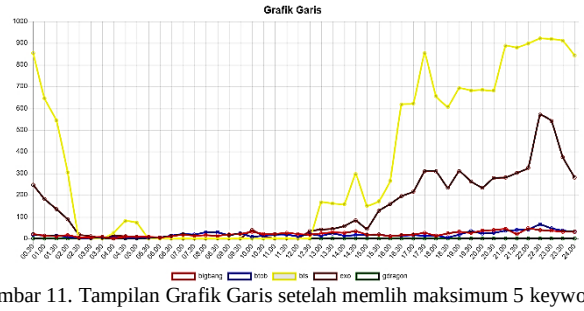
Halaman hasil menampilkan semua informasi yang sesuai dengan *keyword*. Pada bagian kiri atas terdapat tombol untuk kembali ke halaman awal dan tombol untuk menyimpan data yang ada pada *database* dalam format *CSV*. Tampilan halaman hasil dari aplikasi pencarian ini dapat dilihat seperti pada Gambar 10.



Gambar 10. Tampilan halaman hasil.

B. *Hasil Implementasi Grafik*

Pada subbab ini ditampilkan grafik hasil akuisisi data *tweet* dengan sumbu x yaitu waktu *tweet* dalam kelipatan 30menit, sedangkan sumbu y yaitu jumlah *tweet* pada waktu tertentu. Kemudian ditampilkan pilihan *keyword* untuk bisa dipilih dengan maksimal 5 *keyword*. Selanjutnya setelah memilih *keyword*, ditampilkan hasil grafiknya sehingga bisa terlihat perbedaan grafik antara *keyword* yang satu dengan yang lain seperti pada Gambar 11.



Gambar 11. Tampilan Grafik Garis setelah memlih maksimum 5 keyword.

C. *Hasil Clustering dengan WEKA*

Pada Studi ini *Clustering* K-Means dengan Menggunakan WEKA menunjukkan pada iterasi ke-10 perhitungan berhenti sama halnya dengan perhitungan secara manual dengan menggunakan Microsoft Excel. Yang perlu diperhatikan bahwa algoritma K-means menghasilkan hasil yang berbeda-beda karena adanya proses randomisasi menentukan titik awal pusat cluster pada algoritma ini [5]. Pada Gambar 12 ditunjukkan hasil *clustering* dengan menggunakan WEKA.

```
kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 501.170889117187

Initial starting points (random):

Cluster 0: 1340
Cluster 1: 505
Cluster 2: 1033

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
             (68392.0) (35652.0)  (8401.0) (24339.0)
=================================================
time         1004.3145 1270.5107  148.5395  909.7723




Time taken to build model (full training data) : 0.35 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      35652 ( 52%)
1       8401 ( 12%)
2      24339 ( 36%)
```

Gambar 12. Hasil Clustering dengan WEKA.

V. KESIMPULAN

Berdasarkan rangkaian proses yang dilakukan seperti yang telah, maka dapat diambil kesimpulan sebagai berikut :

1. Langkah-langkah dalam mengakuisisi data pada Twitter terdiri dari beberapa tahapan. Tahapan pertama adalah membuat aplikasi pada Twitter dengan memanfaatkan Twitter *API*, dengan login sebagai pengguna Twitter. Setelah itu didapat Consumer Key, Consumer Secret, Access Token, dan Access Secret Token yang berfungsi sebagai kunci untuk bisa mengakses Twitter *API*. Kemudian untuk memastikan keselamatan dan keamanan pada platform Twitter, pengguna yang menggunakan Sign In dengan Twitter harus secara eksplisit mendeklarasikan callback URL. Setelah itu ekstrasi fitur pada Twitter dapat dilakukan.
2. Dari data *tweet* sebanyak 68.392 yang berlokasi pada latitude (-4.640003) dan longitude (109.866141) dengan radius 590KM, mayoritas orang-orang yang melakukan *tweet* K-Pop dan K-Drama dengan 87 *keyword* tertentu membuat *tweet* pada jam 21.00 – 01.00.
3. Hasil penghitungan *clustering* secara manual (menggunakan Excel) dibanding dengan penghitungan *clustering* WEKA mirip meskipun tidak 100% sama. Yang perlu diperhatikan bahwa algoritma K-means menghasilkan hasil yang berbeda karena adanya proses randomisasi pada algoritma K-Means.

DAFTAR PUSTAKA

[1] We Are Social, “Internet User,” *wearesocial.com*, 2018. [Online]. Available: https://wearesocial.com/blog/2018/01/global-digital-report-2018.

[2] Situmorang and et al, *Analisis Data: untuk riset manajemen dan bisnis*. Medan: USU Press, 2010.

[3] Beevolve, “An Exhaustive Study of Twitter Users Across the World,” *www.beevolve.com*, 2012. .

[4] N. Kaur and et al, “Efficient K-Means Clustering Algorithm using Ranking Method in Data Mining,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 3, 2012.

[5] B. Santosa, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.

[6] S. Defiyanti and M. Jajuli, “Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining,” in *Konferensi Nasional Informatika (KNIF)*, 2015.

[7] J. E. Sembodo and et al, “Data Crawling Automatis Pada Twitter,” in *Indonesia Symposium on Computing*, 2016.

[8] D. P. Arumdani, “Pengembangan Sistem Akuisisi Data Twitter Berbasis Web Menggunakan Twitter Streaming API,” Institut Pertanian Bogor, 2016.

[9] D. S. Indraloka and B. Santosa, “Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee,” *J. Sains Dan Seni ITS*, vol. 6, no. 2, 2017.

[10] JSON, “Pengenalan JSON,” *www.json.org*, 2017. [Online]. Available: https://www.json.org/json-id.html.

[11] I. Arın and et al, “I-TWEC: Interactive clustering tool for Twitter,” *Expert Syst. Appl.*, vol. 96, pp. 1–13, 2018.

[12] D. Sayce, “Number of tweets per day,” 2016. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/.

[13] “OpenCL,” *Wikipedia*. Apr-2017.

[14] 3scale, “What is an API,” *www.3scale.net*, 2016. [Online]. Available: https://www.3scale.net/wp-content/uploads/2012/06/What-is-an-API-1.0.pdf.

[15] TulisKode, “Konsep Application Programming Interface,” *www.tuliskode.com*, 2016. [Online]. Available: https://www.tuliskode.com/mengenal-konsep-application-programming-interface-API/.

[16] O. Somantri and et al, “Metode K-Means untuk Optimasi Klasifikasi Tema Makalah Mahasiswa Menggunakan Support Vector Machine (SVM),” *Sci. J. Informatics*, vol. 3, no. 1, 2016.

[17] S. S. Khan and A. Ahmad, “Cluster center initialization algorithm for K-modes clustering,” *Expert Syst. Appl.*, vol. 40, 2013.