Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Penerapan Metode *K-Means Clustering* Dalam Menganalisis Sentimen Masyarakat Terhadap *K-Popers* Pada *Twitter*

Miranti Alysha Zulia Larasati[1], Nurul Anisa Sri Winarsih[2*], Muhammad Syaifur Rohman[3], Galuh Wilujeng Saraswati[4]
Program Studi Teknik Informatika, Universitas Dian Nuswantoro
Jl. Imam Bonjol 207 Semarang, Indonesia
[*]e-mail *Corresponding Author:* nurulanisasw@dsn.dinus.ac.id

Abstract
The evolution of Twitter as a platform loved by the general public in Indonesia is evidenced by statistical data that shows that Indonesia is ranked 7th in the world and has a large number of users, reaching 13.2 million. Many users have expressed their opinions on Twitter. This includes expressions with hate speech to bullying. From this research, an analysis was carried out on public satisfaction with K-Pop to get a benchmark for how far people know the existence of K-Pop in Indonesia. This research was conducted using the K-Means Clustering algorithm method to group (*positive, neutral and negative*) sentiments from datasets taken from *Twitter*. The dataset used consists of 1000 data retrieved according to the results of the polarity of the tweet. Based on the test results, it got a negative sentiment value of 15.09%, neutral 51.75%, and positive 33.15%. With the evaluation level using the silhouette coefficient method, which is 0.687974 which means it has good structural results.
Keywords*: Data Mining, K-means Clustering, Silhouette coefficient, K-pop*

Abstrak
Evolusi *Twitter* sebagai *platform* yang digemari masyarakat umum di Indonesia dibuktikan dengan data statistik yang menunjukkan bahwa Indonesia menempati peringkat ke-7 dunia dan memiliki jumlah pengguna yang besar yaitu mencapai 13,2 juta. Banyak pengguna yang mengutarakan pendapat di *Twitter*. Ini termasuk ekspresi dengan ujaran kebencian hingga perundungan. Dari penelitian tersebut dilakukan analisis tentang kepuasan masyarakat terhadap *K-Pop* untuk mendapatkan tolok ukur seberapa jauh masyarakat mengetahui eksistensi *K-Pop* di Indonesia. Penelitian ini dilakukan dengan menggunakan algoritme *K-Means Clustering* untuk mengelompokkan sentimen positif, netral dan negatif dari dataset yang diambil dari *twitter*. Dataset yang digunakan terdiri dari 1000 data yang diambil sesuai hasil polaritas *tweet*. Berdasarkan hasil pengujian mendapatkan nilai sentimen negatif sebanyak 15,09%, netral 51,75%, dan positif 33,15%. Dengan tingkat evaluasi menggunakan metode *Silhouette Coefficient* yaitu sebesar 0.687974 yang berarti memiliki hasil struktur yang baik.
Kata kunci: *Data Mining, K-means Clustering , Silhouette coefficient, K-pop*

1. Pendahuluan

Popularitas *Korean Wave* atau lebih dikenal dengan *Hallyu* pertama kali berkembang hanya di negara-negara Asia Timur dan kemudian menjadi populer di seluruh dunia hingga Indonesia (Jin, 2016). Perkembangan budaya Korea saat ini sangat populer di kalangan pemuda dan dewasa, baik perempuan maupun laki-laki[1], [2]. Kata *Hallyu* diadopsi pada tahun 1999 oleh Kementerian Kebudayaan dan Pariwisata Korea Selatan dalam perencanaan, produksi dan distribusi yang digunakan oleh CD musik musisi Korea di negara tetangga atau dalam bahasa Inggris *Korean Pop Music.* Dalam bahasa Cina juga disebut sebagai *Hallyu – Song of Korea* (Musik Korea)[3]. Istilah *hallyu* telah mendapatkan popularitas yang luas sejak surat kabar China melaporkan keberhasilan penyanyi Korea di China.

Korean wave atau gelombang *hallyu* yang menyeluruh di berbagai macam belahan dunia tidak terkecuali di wilayah Indonesia dengan tingkat perkembangannya sangat pesat juga tidak luput oleh penggunaan adanya internet termasuk juga social media. Akun media sosial penggemar *K-pop* digunakan untuk mengakses berbagai informasi tentang idola mereka [3], [4]. Menurut survei kumparan, 56 persen dari penggemar *K-pop* menghabiskan 1-5 jam menjelajahi media sosial untuk menemukan semua informasi yang mereka butuhkan tentang idola mereka [3].

Dan saat ini perkembangan teknologi komunikasi juga sangat mendukung dengan menawarkan sesuatu yang positif untuk memudahkan kita menerima informasi atau mengirimkan informasi melalui *Website* atau jejaring sosial. Tidak hanya hal positif namun dampak negatif pun terkadang dapat terjadi dalam perkembangan teknologi. Misalnya, Ujaran kebencian, pencurian data, penipuan transaksi *online* dan berita bohong atau *hoax* [3], [5]. Hal ini menyebabkan muncul berbagai pendapat oleh masyarakat dan memunculkan pro dan kontra mengenai komentar masyarakat tentang eksistensi *K-pop* di media sosial. Media sosial yang digunakan dalam penelitian ini adalah *twitter*. Berdasarkan data dari *website* (https://databoks.katadata.co.id/) Twitter berada di peringkat 5 platform media sosial yang sering digunakan pada tahun 2020 [6],[7].

Konsep Data mining sangat cocok diterapkan untuk mengetahui pola dari suatu data sentimen masyarakat yang dilakukan didalam media sosial *Twitter*. Algoritma *unsupervised* adalah algoritma *Machine Learning* (ML) yang digunakan untuk kumpulan data yang tidak berlabel, yaitu yang tidak memiliki variabel keluaran (*output*)[8]. Algoritma *unsupervised* memfasilitasi analisis kumpulan data, sehingga membantu menghasilkan informasi dari data yang tidak berlabel. Kemajuan terbaru dalam pembelajaran hierarkis, algoritma pengelompokan, analisis faktor, model laten, dan deteksi *outlier*, telah membantu secara signifikan dalam teknik *unsupervised machine learning*. Metode *Clustering* merupakan metode analisa data yang dapat digunakan dalam memecahkan masalah dalam suatu pengelompokan data. Salah satu metode yang ada di dalam metode *clustering* adalah metode *K-mens* [9], [10]. Metode *K-means* merupakan suatu metode yang dapat melakukan pengelompokan data dalam jumlah yang cukup besar dengan perhitungan waktu yang relatif cepat dan efisien. Dalam artikel ini diuji penggunaan algoritma *K-Means* dalam mengelompokkan sentimen masyarakat terhadap eksistensi *K-Pop* pada media sosial *twitter*.

Penelitian ini menganalisis sentimen masyarakat Indonesia terhadap eksistensi *K-Pop* pada media sosial *twitter*. Kinerja dari sebuah algoritma klasifikasi dipengaruhi dari jenis data dan fiturnya, maka dari itu data set berupa teks yang akan diolah harus melalui tahapan *Text preprocessing* seperti *case folding, stemming, tokenizing, Text Normalization* serta *stopwords*, lalu setelah itu data akan masuk tahapan selanjutnya yaitu tahapan klasifikasi menggunakan algoritma *K-Means* dan diuji dengan perhitungan *Silhouette Coefficient* untuk mendapatkan nilai akurasi yang sesuai dengan harapan sehingga dapat mengklasifikasikan data untuk mendapatkan hasil kesimpulan[11][12].

2. Tinjauan Pustaka

Beberapa paparan terdahulu dengan metode yang sama yaitu Implementasi dari pendekatan metode *K-Means* untuk mengetahui kecenderungan opini masyarakat terhadap pemilu dalam media sosial *twitter*. Pada penelitian ini dilakukan untuk mengetahui kecenderungan opini masyarakat terhadap pemilu apakah termasuk kedalam sentimen positif atau negative [6].

Analisis sentimen sebelumnya juga menggunakan metode Algoritma *K-Means* untuk menganalisis sentimen dari ulasan film dengan memvalidasi keakuratan metode *K-Means* untuk data ulasan film. Penelitian yang dilakukan Setyo Budi dalam penelitiannya yang membahas tentang *text mining* guna menganalisis opini ulasan film dengan metode algoritma *K-means* menghasilkan nilai akurasi sebanyak 57,83% [8].

Dalam penelitian yang dilakukan Rahmawati, Marjuni dan Zeniarja [13] mengenai Analisis Sentimen Publik pada media social *Twitter* mengenai pelaksanaan pilkada serentak menggunakan algoritma *K-Means* dan *Support Vector Machine*. Penelitian ini bertujuan untuk mengetahui respon masyarakat pada media sosial *Twitter* tentang kelangsungan pilkada. Penelitian ini menggunakan 3000 tweet Bahasa Indonesia yang digunakan untuk dataset dan membagi data kedalam 2 kategori yaitu *Cluster* 1 sebagai kelompok *Tweet* positif dan *Cluster* 2 sebagai kelompok *Tweet* negatif.

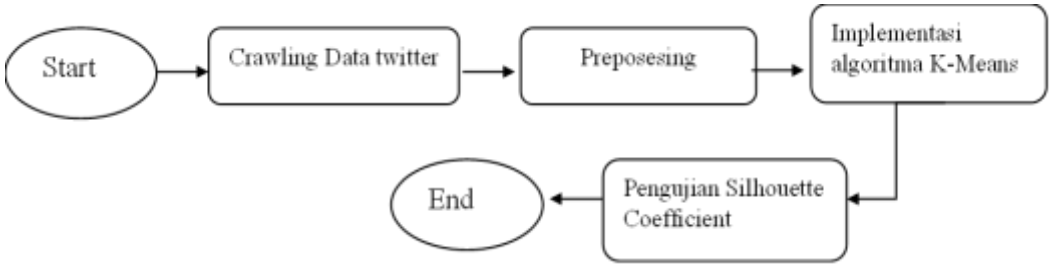
Pada tahun 2020 penelitian yang dilakukan Faesal, Muslim, Ruger dan Kusrini [14] yang berjudul Sentimen Analisis pada Data *Tweet* Pengguna *Twitter* Terhadap Produk Penjualan Toko *Online* Menggunakan Metode *K-Means*. Penelitian ini bertujuan untuk mengetahui bagaimana sentimen publik terhadap keputusan konsumen terhadap produk penjualan sehingga menjadi peluang bagi produsen dalam mempromosikan dan memasarkan produknya kepada konsumen. Penelitian ini melakukan pengumpulan data sebanyak 1.000.000 tweet data *tweet* pada media sosial *Twitter* yang berkaitan dengan penjualan di Tokopedia. Dengan pengujian berdasarkan kata *tweet* diperoleh tingkat akurasi sebesar 92.80 %.

Yan Watequlis Syaifudin, Rizki Andi Irawan "Analisis *Clustering* Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means" penelitian dilakukan pada tahun 2018. Algoritma yang digunakan dalam klasifikasi adalah *K-Means* dan *Support vector machine*. Data opini diperoleh dari jejaring sosial *Twitter* dalam Bahasa Indonesia dengan topik suatu pantai. Klasifikasi opini diperlukan untuk memudahkan pengguna dalam melihat opini positif, negatif, ataupun netral [15].

Aditama, Pratama, Wiwaha, dan Rakhmawati [16] melakukan penelitian pada tahun 2020 tentang "Analisis Klasifikasi Sentimen Pengguna Media Sosial *Twitter* Terhadap Pengadaan Vaksin COVID-19". Penelitian ini bertujuan untuk mengetahui bagaimana sentimen publik terhadap sebuah masalah atau objek, apakah cenderung beropini negatif atau positif. Penelitian ini melakukan pengumpulan data dengan melakukan *crawling twitter* dan menghasilkan 1000 tweet untuk dataset. Dan menghasilkan persentase opini masyarakat terhadap vaksin *Corona* yaitu 48% positif, 29% netral, dan 23% negative.

Perbedaan dari penelitian-penelitian yang sudah dilakukan sebelumnya, yaitu penelitian mengenai Analisis Sentimen Seputar opini public terhadap *K-Popers* pada media sosial *Twitter* menggunakan metode *K-Means* menggunakan Bahasa pemrograman Python untuk proses pengolahannya. Penelitian ini bertujuan untuk mengimplementasikan algoritma *K-Means* terhadap dataset seputar opini masyarakat mengenai sentimen pada *K-pop*, dengan dataset berjumlah 1000. Penelitian ini diharapkan dapat dijadikan sebagai salah satu parameter untuk menjadi penunjang bagi masyarakat dalam mengevaluasi pengambilan keputusan dan kebijakan di masa yang akan datang dan memberikan gambaran hasil pelabelan sentimen.

3 Metodologi



Gambar 1. Peta Konsep Penelitian

Penelitian ini dimulai dari pengumpulan dataset, pengolahan sentimen masyarakat pada media sosial *Twitter* menggunakan bahasa pemrograman *phyton*. Proses pada penelitian ini juga bersifat ekperimental untuk menguji dan mengevaluasi keakuratan sentimen positif, netral dan negatif pada *tweet* masyarakat terhadap eksistensi *K-Pop* di Indonesia.

3.1 Algoritma *K-Means*

K-Means merupakan sebuah algoritma clustering pada data mining untuk dapat menghasilkan kelompok dari data yang jumlahnya banyak dengan metode partisi yang berbasis titik dengan waktu komputasi yang cepat dan efisien [17]. *K-Means* merupakan salah satu dari metode pengelompokkan data non hierarki (sekatan) yang dapat mempartisi data kedalam bentuk dua kelompok ataupun lebih [10] Data-data diseleksi menjadi beberapa bagian dengan kriteria yang sudah ditetapkan kemudian digabungkan jadi satu dalam cluster. Algoritma *K-Means* disajikan sebagai berikut [18]:

1) menentukan jumlah kelompok atau *cluster* (k), lalu pilih pusat grup secara acak.
2) menghitung jaraknya dari setiap data ke pusat *cluster*.

Di dalam proses clustering, dapat diawali dengan mengenali data yang dikelompokkan, menggunakan rumus Euclidean Distance.

Rumus *Euclidean Distance*:

$$M_{(a,b)} = \sqrt{(X_{1a} - X_{1b})^2 + (X_{1a} - X_{1b})^2 + \ldots + (X_{na} - X_{nb})^2} \quad \ldots\ldots\ldots\ldots (1)$$

Keterangan:
M (a, b) = jarak data dari a ke pusat b
Xka = Data dari a pada atribut data ke k
Xkb = Data dari b pada atribut data ke k

3) mengelompokkan data ke dalam kelompok-kelompok dengan jarak terdekat, kemudian hitung pusat kelompok yang baru
4) hitung pusat grup baru ditambahkan dengan anggota kelas, dengan menggunakan cara mencari mean (rata-rata) dari objek dalam kelas yang lebih spesifik.
5) kemudian ulangi langkah ke-2, ketiga dan seterusnya hingga tidak ada data yang dipindahkan ke cluster lain.

3.2 Silhouette Coefficient

Untuk mencapai akurasi yang akurat dari hasil proses *clustering* dapat diketahui melalui perhitungan nilai *silhouette coefficient*. Pengujian dilakukan setelah mencapai konvergensi 0 di mana hasil pengelompokan terakhir sama dengan pengelompokan sebelumnya. Dengan kata lain, tidak ada data yang berpindah klaster[19]. Pengujian dihitung menggunakan persamaan *Silhouette coefficient*. Berikut tahapan untuk mengkalkulasi nilai *silhouette coefficient*:

1) Mengestimasi rata-rata jarak objek ke-i dengan seluruh titik yang terdapat didalam satu cluster. Dapat ditulis dengan persamaan berikut:

$$a(\mathrm{i}) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq 1} \mathrm{d}(\mathrm{i}, j) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

dimana:
A = banyaknya data dalam kelas A

2) Menghitung suatu nilai b(i) Ini adalah jumlah minimum jarak rata-rata antara data ke-i dan semua data di kelas yang berbeda. Menghitung jarak rata-rata antara data Ke-i dan data dalam *cluster* lain (misal C) ditulis sebagai berikut:

$$\mathrm{d}(\mathrm{i}, C) = \frac{1}{|C|} \sum_{j \in C} \mathrm{d}(\mathrm{i}, j) \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

dengan keterangan dimana C = banyaknya data dalam kelas C
setelah menghitung d(I, C) untuk semua cluster atau kelas C ≠ A, kemudian memilih nilai jarak paling kecil sebagai nilai b(i).

$$b(\mathrm{i}) = \min_{C \neq A} \mathrm{d}(\mathrm{i}, j) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

3) Menghitung sebuah hasil *silhouette coefficient* bisa ditulis dengan menggunakan rumus sebagai berikut:

$$s(\mathrm{i}) = \frac{b(\mathrm{i}) - a(\mathrm{i})}{max\{a(\mathrm{i}) - b(\mathrm{i})\}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Penafsiran nilai *silhouette coefficient* dapat ditunjukkan dalam tabel berikut:

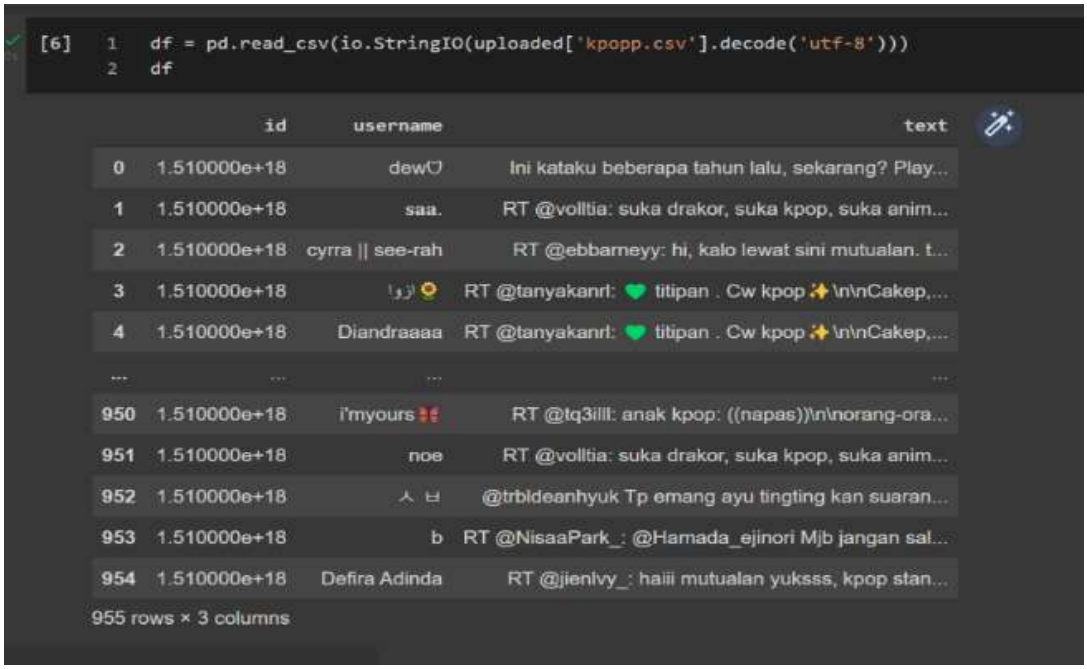
Tabel 1. Interpretasi Nilai *Sillhoette Coefficient*

Nilai *Silhoette Coefficient*	Interpretasi
0.71 – 1.00	Hasil susunan yang kuat
0.51 – 0.70	Hasil susunan yang baik
0.26 − 0.50	Hasil susunan yang lemah
≤ 0.25	Data tidak tersusun

4 Hasil dan Pembahasan

4.1 pengumpulan data

Data yang digunakan dalam sistem ini berupa *tweet* dengan jumlah kurang lebih 1000 data yang diambil dari *twitter* dengan pencarian kata kunci tertentu. *Crawling* data dilakukan kurang lebih selama 1 bulan. Hasil *crawling* data dapat dilihat pada gambar dibawah ini.



```
[6]  1  df = pd.read_csv(io.StringIO(uploaded['kpopp.csv'].decode('utf-8')))
     2  df
```

	id	username	text
0	1.510000e+18	dewひ	Ini kataku beberapa tahun lalu, sekarang? Play...
1	1.510000e+18	saa.	RT @volltia: suka drakor, suka kpop, suka anim...
2	1.510000e+18	cyrra || see-rah	RT @ebbarneyy: hi, kalo lewat sini mutualan. t...
3	1.510000e+18	زوا🌻	RT @tanyakanrl: 💚 titipan . Cw kpop✨\n\nCakep,...
4	1.510000e+18	Diandraaaa	RT @tanyakanrl: 💚 titipan . Cw kpop✨\n\nCakep,...
...	...	...	...
950	1.510000e+18	i'myours🎀	RT @tq3ill: anak kpop: ((napas))\n\norang-ora...
951	1.510000e+18	noe	RT @volltia: suka drakor, suka kpop, suka anim...
952	1.510000e+18	ㅅㅂ	@trbldeanhyuk Tp emang ayu tingting kan suaran...
953	1.510000e+18	b	RT @NisaaPark_: @Hamada_ejinori Mjb jangan sal...
954	1.510000e+18	Defira Adinda	RT @jienlvy_: haiii mutualan yuksss, kpop stan...

955 rows × 3 columns

Gambar 2. Sample Data Hasil *Crawling*

4.2 Preprocessing

Setelah mendapatkan hasil *crawling data* dan dijadikan dataset tahap selanjutnya yaitu *text preprocessing.* Dengan proses mengubah data mentah yang tidak terstruktur menjadi data *tweet* yang baik dan siap untuk diolah dengan jumlah *cluster* yang digunakan sebanyak 3 *cluster* yaitu kelas C0=Positif, C1=Netral, dan C2=Negatif.

Tabel 2.Contoh Hasil *Tweet*

Tweet atau Data mentah	Preprocessing
RT @MyQueenJiharu: Sekarang banyak orang kehilangan adab karena kpop, ngetik jahat/ gomong jht bisa tapi etika minta maaf ngk ada.	orang hilang adab kpop ngetik jahat gomong jht etika maaf ngk
astaga diam itu tiketnya MAHAL BGT tunggu artis2 kpop ku aja deh	astaga diam tiket mahal bgt tunggu artis kpop ku aja deh
Lebih mahal dari konser kpop hiks https://t.co/5wOCBsd6K7	mahal konser kpop hiks
Aku sebagai anak sastra yang suka kpop, drakor dan musik indie hanya bisa tertawaðŸ˜, https://t.co/AQUxsZSGFx	anak sastra suka kpop drakor musik indie tertawa
RT @volltia: suka drakor, suka kpop, suka anime, suka film horror, suka makan, suka uang, suka belanja di shopee, dan suka kamu juga, kamuâ€¦	suka drakor suka kpop suka anime suka film horror suka makan suka uang suka belanja shopee suka
Sayang bgt kurang kesal KPop https://t.co/ClwBApoPr2	sayang banget kurang kesal kpop
@natanattda Hooh kukira grup kpop baruuu ternyata jpop, mana katanya hikaru abis dari kepler bakal masuk xg rumornya keren bgt deh	hooh kira grup kpop baruuu jpop hikaru habis kepler masuk rumor keren banget
agensi agensi kpop tuh udah tau kalo merch yg paling banyak diminati adalah photocardnya jadi apa apa ada pcnya skg WKWKWKWKðŸ˜-ðŸ‘•	agens agens kpop tuh udah tau kalo merch yg mati photocardnya pcnya skg wkwkwkwk

Tweet atau Data mentah	Preprocessing
@ctxvrll27 Yang kpop experiment ngerate fandom itu, akunnya baru sih, tapi kontennya asdfghjkl wkwkkw	akun kpop experiment loh ngerate fandom kwkkw
pengalaman hidupku yang paling berkesan adalah aku kenal kpop, mengubah hidup banget cuy https://t.co/vc4g98MAKl	alam hidup kesan kenal kpop ubah hidup banget cuy

4.3 Implementasi menggunakan K-Means

1) Menentukan banyak *cluster* yang ingin dibentuk.
   Jumlah *cluster* yang ingin dibentuk dan ditetapkan pada artikel ini, yaitu 3 cluster (negative, neutral, positif).
2) Proses perhitungan K-Means yang dilakukan adalah sebagai berikut:
   Iterasi ke-1, menentukan titik pusat awal atau *centroid* secara acak

Tabel 3. *Tweet* yang Menjadi *Centroid*

Data ke-	Tweet	Cluster
642	jatuh cinta irl ngebuat jarang hype kpop anjir	C0
87	closefriends ig gw kaya lelah liat cowo kpop tengger	C1
302	bener yg bikin gue suka kpop multi tpi klo selesai gue selesai kpopan	C2

3) Menghitung jarak terdekat setiap data dengan *centroid*.
   Hitung masing-masing *cluster* ke setiap titik pusat dengan menggunakan persamaan *Euclidean distance*.

$$M_{(a,b)} = \sqrt{(X_{1a} - X_{1b})^2 + \ldots\ldots + (X_{na} - X_{nb})^2}$$

$$M11 = \sqrt{(1-1)^2 + (3-0)^2} = \sqrt{9} = 3$$
$$M12 = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2} = 1{,}41$$
$$M13 = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2} = 1{,}41$$
$$\ldots\ldots$$
$$M110 = \sqrt{(0-1)^2 + (0-0)^2} = \sqrt{1} = 1$$
$$M21 = \sqrt{(1-0)^2 + (3-0)^2} = \sqrt{10} = 3{,}16$$
$$M22 = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1} = 1$$
$$M23 = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1} = 1$$
$$\ldots\ldots$$
$$M210 = \sqrt{(0-0)^2 + (0-0)^2} = \sqrt{0} = 0$$
$$M31 = \sqrt{(1-0)^2 + (3-1)^2} = \sqrt{5} = 2{,}23$$
$$M32 = \sqrt{(0-0)^2 + (1-1)^2} = \sqrt{0} = 0$$
$$M33 = \sqrt{(0-0)^2 + (1-1)^2} = \sqrt{0} = 0$$
$$\ldots\ldots$$
$$M310 = \sqrt{(0-0)^2 + (0-1)^2} = \sqrt{2} = 1{,}41$$

Tabel 4. Hasil Perhitungan *K-Means* pada Iterasi 1

Data	Jarak Ke C0	Jarak Ke C1	Jarak Ke C2	Jarak Terdekat
1	3	3,16	2	C2
2	1,41	1	1	C1
3	1,41	1	1	C1
4	1	2	1,41	C0
5	7	8	7,07	C0
6	1,41	2,23	1	C2
7	1	2	1,41	C0
8	1,41	1	1	C1
9	1	0	1,41	C1
10	1	0	1,41	C1

Dari tabel 4 pada interasi 1, terdapat nilai dari keanggotaan dari masing masing klister, 3 data yang tergabung dalam pusat claster 0 (C0), terdapat 5 data yang tergabung dalam C1, terdapat 2 data yang tergabung dalam C2. Keanggotaan cluster dapat dihitung dengan membandingkan jarak terkecil antara ketiga *cluster* tersebut. Kemudian langkah selanjutnya yaitu dengan membentuk titik pusat baru untuk melakukan iterasi 2. Perhitungan tersebut dilakukan dengan rumus dan cara yang sama dengan iterasi 1. Berikut hasil dari interasi 2 dengan titik pusat yang baru.

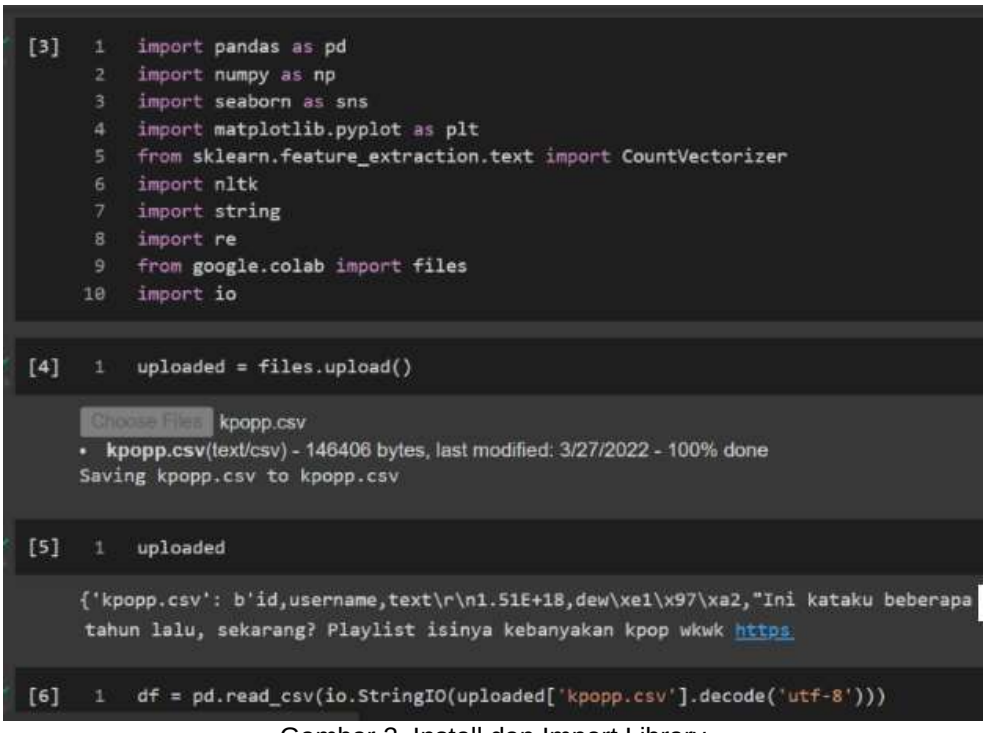
Tabel 5. Hasil Perhitungan *K-Means* pada Iterasi 2

Data	Jarak Ke C0	Jarak Ke C1	Jarak Ke C2	Jarak Terdekat
1	4,24	2,6	0,5	C2
2	4,12	0,4	1,80	C1
3	4,12	0,4	1,80	C1
4	2	2,08	2,69	C0
5	4	8,02	7,43	C0
6	3,60	1,72	0,5	C2
7	2	2,08	2,69	C0
8	4,12	0,4	1,80	C1
9	4	0,6	2,69	C1
10	4	0,6	2,69	C1

Perhitungan akan di ulang terus menerus sampai dengan rata-rata iterasi sebelumnya dengan rata-rata ietrasi selanjutnya tidak mengalami perubahan, jika sudah tidak mengalami perubahan dari iterasi 1 maka didapatkan hasil klaster pengelompokan data tweet.

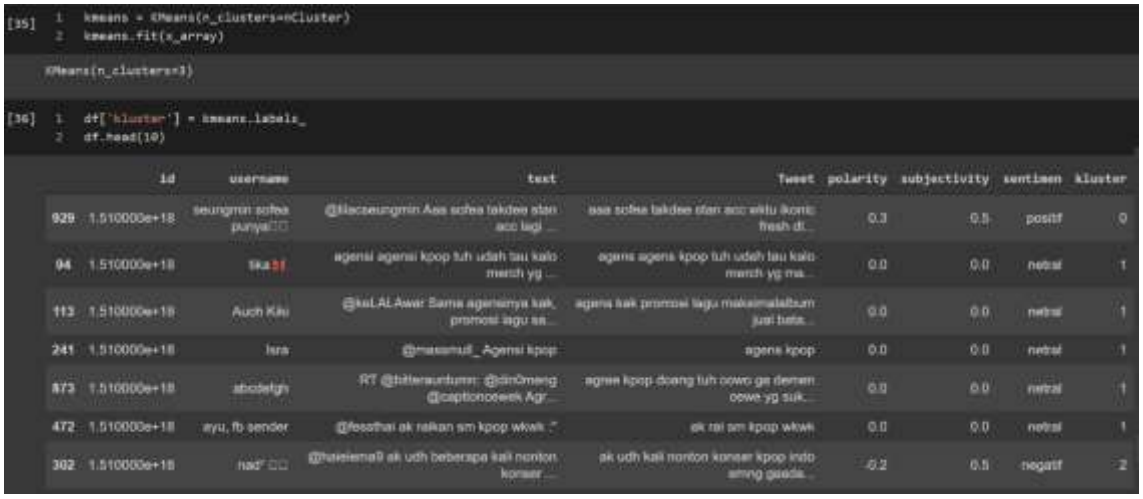
4.4 Implementasi *Google Colaboratory*

Hal pertama yang harus dilakukan sebellum menjalankan proses *preprocessing* pada *google colaboratory* adalah menyiapkan library yang diperlukan dan juga data *tweet*.



Gambar 3. Install dan Import Library

Setalah selesai dengan pengaturan tentang proses *preprocessing* langkah selanjutnya adalah proses *clustering K-means*. Hasil perhitungan dengan *google colaboratory* didapatkan yaitu data positif (C0) sebanyak 33,15%, tdata netral (C1) sebanyak 51,75% dan data negatif (C2) sebanyak 15,09% data.



```
[35] 1 kmeans = KMeans(n_clusters=nCluster)
     2 kmeans.fit(x_array)

KMeans(n_clusters=3)

[36] 1 df['kluster'] = kmeans.labels_
     2 df.head(10)
```

	id	username	text	Tweet	polarity	subjectivity	sentimen	kluster
929	1.510000e+18	seungmin sofea punya□□	@lilacseungmin Aaa sofea takdee stan acc lagi ...	aaa sofea takdee stan acc wktu ikonic fresh di...	0.3	0.5	positif	0
94	1.510000e+18	Ska🌻	agensi agensi kpop tuh udah tau kalo merch yg ...	agens agens kpop tuh udah tau kalo merch yg ma...	0.0	0.0	netral	1
113	1.510000e+18	Auch Kiki	@kuLALAwar Sama agensinya kak, promosi lagu sa...	agens kak promosi lagu makasimalkan just bata...	0.0	0.0	netral	1
241	1.510000e+18	lara	@masamull_ Agensi kpop	agens kpop	0.0	0.0	netral	1
873	1.510000e+18	abodelgh	RT @bitterandumm: @dinOmeng @captioncewek Agr...	agree kpop doang tuh cowo ge demen cewe yg suk...	0.0	0.0	netral	1
472	1.510000e+18	ayu, fb sender	@feasthai ak nakan sm kpop wkwk :"	ak nai sm kpop wkwk	0.0	0.0	netral	1
302	1.510000e+18	nad' □□	@haeieima9 ak udh beberapa kali nonton konser...	ak udh kali nonton konser kpop indo emng gasda...	-0.2	0.5	negatif	2

Gambar 4. Hasil Pengelompokan *Cluster*

4.5 Pengujian dengan *Silhouette Coefficient*

Untuk melihat kemiripan data yang tergabung ke dalam satu kelompok, maka perlu dilakukan uji homogenitas. Pengujian ini dilakukan dengan menggunakan metode *silhouette coefficient*, yang akan menghasilkan nilai -1 jika hasil dari klastering buruk dan mendekati 1 jika hasil dari klastering baik. Hasil perhitungan nilai (ai), b(i), dan *Silhouette coefficient* (s) masing-masing data ditunjukkan dalam tabel berikut:

Tabel 6. hasil rata-rata *Sillhoette Coefficient*

No	Data Tweet	Cluster	a(i)	b(i)	S(i)
1	4	1	3	2.1416	-0.286
2	5	1	6	8.0373	0.2534
3	7	1	3	2.1416	-0.286
4	2	2	0.5	1.8251	0.3171
5	3	2	0.5	1.8251	0.3171
6	8	2	0.5	1.8251	0.3171
7	9	2	0.75	2.6991	0.4872
8	10	2	0.75	2.6991	0.4872
9	1	3	1	2.6065	0.3457
10	6	3	1	1.7429	0.1896
		Rata-rata			0.2142

Jika dilihat dari hasil contoh pengujian menggunakan *silhouette coefficient* pada Tabel 6, terdapat 2 data yang bernilai negatif yaitu data *tweet* 1 dan 3 yang artinya data tersebut seharusnya tidak terletak pada kelompok atau klaster tersebut atau digolongkan dengan lemah pada klaster tersebut. Sedangkan rata-rata nilai *silhouette coefficient* pada contoh pengujian 10 data tersebut adalah 0.214271 yang artinya menurut tabel 2 adalah struktur yang dihasilkan tidak terstruktur. Hal ini dapat terjadi disebabkan karena penentuan letak *centroid* awal yg tak optimal sehingga menghasilkan hasil *cluster* yang kurang baik serta kondisi tersebut juga ditentukan oleh metode perthitungan jarak yang digunakan. Dan pengujian pada seluruh data menghasilkan nilai *Silhoute Index* dengan *Euclidian Distance* 0.687974 sehingga menurut pada tabel 2 menghasilkan hasil susunan yang baik.

Algoritma *K-means* sering digunakan *untuk text clustering*. Seperti halnya pada penelitian Imam dan Ajib, mereka melakukan *text clustering* pada tweet Pilpres 2019 [6] serta penelitian Rahmawati dan tim tentang sentimen publik pilkada serentak [13]. *K-means* juga

digunakan pada bidang hiburan seperi sentimen review *film* [8], sentiment Wisata Pantai [15], bidang perekonomian seperti sentimen produk penjualan toko *online* [14], hingga bidang kesehatan yaitu sentiment pengadaan Vaksin COVID-19 [16]. Setelah proses *clustering*, data diuji dengan perhitungan *Silhouette Coefficient* seperti yang diteliti oleh Hidayati [19]. Berdasarkan hasil penelitian ini, *text clustering* menggunakan *K-means* dengan *silhouette coefficient* menguatkan penelitian [6], [8], [13-16]

5. Simpulan

Berdasarkan hasil klasifikasi data sentimen *twitter* dan hasil pengujian terdahap kalsifikasi dengan menggunakan metode *K-Means Clustering*, maka dapat ditarik kesimpulan bahwa metode *K-Means Clustering* baik dalam melakukan klasifikasi sentimen masyarakat pada media sosial *twitter* dengan menghasilkan nilai *Silhoute Index* 0.687974 yang artinya dataset tersebut digolongkan pada struktur yang baik. Hal tersebut dapat dipengaruhi oleh data uji dan data latih yang kualitas sangat baik, sehingga saat dilakukan uji coba terhadap data lain dapat mempengaruhi hasil klasifikasi. Namun dari beberapa penelitian sebelumnya dan hasil klasifikasi dalam penelitian ini maka metode *K-Means Clustering* dapat digunakan untuk melakukan klasifikasi sentimen masyarakat pada media sosial *twitter*. Oleh karena itu disarankan untuk mencoba melakukan klasifikasi dengan data sentimen lain atau data yang sama.

Daftar Referensi

[1] K. Zakiah, D. Widya Putri, N. Nurlimah, D. Mulyana, And Nurhastuti, “Menjadi Korean Di Indonesia: Mekanisme Perubahan Budaya Indonesia-Korea,” *Media Tor*, Vol. 12, No. 1, Pp. 90–101, 2019,

[2] M. Kim, Y. C. Heo, S. C. Choi, And H. W. Park, “Comparative Trends In Global Communication Networks Of #Kpop Tweets,” *Qual. Quant.*, Vol. 48, No. 5, Pp. 2687–2702, 2014, Doi: 10.1007/S11135-013-9918-1.

[3] A. R. Rinata And S. I. Dewi, “Fanatisme Penggemar Kpop Dalam Bermedia Sosial Di Instagram,” *Interak. J. Ilmu Komun.*, Vol. 8, No. 2, Pp. 13–21, 2019, Doi: 10.14710/Interaksi.8.2.13-21.

[4] R. S. Tanjung, “Motivasi Dan Perilaku Penggemar Musik Korean Pop Di Medan,” P. 83, 2019, [Online]. Available: Http://Repository.Umsu.Ac.Id/Bitstream/123456789/7289/1/ Motivasi Dan Perilaku Penggemar Musik Korean Pop Di Medan.Pdf

[5] Y. V. Wijaya, A. Erfina, And C. Warman, “Analisis Sentimen Seputar Uu Ite Menggunakan Algoritma Support Vector Machine,” *Progresif J. Ilm. Komput.*, Vol. 17, No. 2, Pp. 1–14, Aug. 2021, Doi: 10.35889/Progresif.V17i2.644.

[6] I. Kurniawan And A. Susanto, “Implementasi Metode K-Means Dan Naïve Bayes Classifier Untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019,” *Eksplora Inform.*, Vol. 9, No. 1, Pp. 1–10, Sep. 2019, Doi: 10.30864/Eksplora.V9i1.237.

[7] N. Puji, S. Endang, And T. Listyorini, “Analisis Sentimen Terhadap Penggemar K-Pop Di Media Sosial Twitter Menggunakan Naive Bayes (Studi Kasus Penggemar Grup BTS”, *Journal Information Engineering and Educational Technology*, Vol.4, No. 2, Pp. 86–89, 2020.

[8] S. Budi, “Text Mining Untuk Analisis Sentimen Review Film,” *Techno.Com*, Vol. 16, No. 1, Pp. 1–8, 2017.

[9] M. W. Goni and Sembiring, "Implementasi K-Means Dalam Pengelompokan Penyebaran COVID-19 di Jawa Barat". *Progresif: Jurnal Ilmiah Komputer*, vol. 17, no. 2, pp. 107-118, 2021.

[10] S. Suhartini, R. Yuliani, G. Gustientiedina, M. H. Adiya, And Y. Desnelita, “Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan,” *J. Nas. Teknol. Dan Sist. Inf.*, Vol. 4, No. 1, Pp. 39–50, 2021, Doi: 10.25077/Teknosi.V5i1.2019.17-24.

[11] I. K. A. Wirayasa And H. Santoso, “Analisis Employee Satisfaction Menggunakan Teknik Clustering Dan Classification Machine Learning,” *Progresif J. Ilm. Komput.*, Vol. 18, No. 1, Pp. 1–10, Jan. 2022, Doi: 10.35889/Progresif.V18i1.766.

[12] K. Fatmawati And A. P. Windarto, “Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi,” *Comput. Eng. Sci. Syst. J.*, Vol. 3, No. 2, Pp. 173–178, 2018, Doi: 10.24114/Cess.V3i2.9661.

[13] A. Rahmawati, A. Marjuni, And J. Zeniarja, “Analisis Sentimen Publik Pada Media Sosial Twitter Terhadap Pelaksanaan Pilkada Serentak Menggunakan Algoritma Support Vector Machine,” *Ccit J.*, Vol. 10, No. 2, Pp. 197–206, 2017, Doi: 10.33050/Ccit.V10i2.539.
[14] A. Faesal, A. Muslim, A. H. Ruger, And K. Kusrini, “Sentimen Analisis Terhadap Komentar Konsumen Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means,” *Matrik J. Manajemen, Tek. Inform. Dan Rekayasa Komput.*, Vol. 19, No. 2, Pp. 207–213, 2020, Doi: 10.30812/Matrik.V19i2.640.
[15] Y. W. Syaifudin And R. A. Irawan, “Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means,” *J. Inform. Polinema*, Vol. 4, No. 3, Pp. 189–194, 2018, Doi: 10.33795/Jip.V4i3.205.
[16] N. A. Rakhmawati, M. I. Aditama, R. I. Pratama, And K. H. U. Wiwaha, “Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin Covid-19,” *J. Inf. Eng. Educ. Technol.*, Vol. 4, No. 2, Pp. 90–92, 2020, Doi: 10.26740/Jieet.V4n2.P90-92.
[17] S. Suhartini And R. Yuliani, “Penerapan Data Mining Untuk Mengcluster Data Penduduk Miskin Menggunakan Algoritma K-Means Di Dusun Bagik Endep Sukamulia Timur,” *Infotek J. Inform. Dan Teknol.*, Vol. 4, No. 1, Pp. 39–50, 2021, Doi: 10.29408/Jit.V4i1.2986.
[8] A. Jananto, "Penerapan Algoritma K-Means Clustering Untuk Perencanaan Kebutuhan Obat di Klinik Citra Medika". *Progresif: Jurnal Ilmiah Komputer*, vol. 18, no. 1, pp. 69-76, 2022.
[19] R. Hidayati *Et Al.*, “Analisis Silhouette Coefficient Pada 6 Perhitungan Jarak K-Means Clustering,” Vol. 20, No. 2, Pp. 186–197, 2021.