Implementasi Text Mining Terhadap Opini Netizen Mengenai PSBB Covid-19 Menggunakan K-Means Clustering

Muh. Syahwal[1], Magfirah Ahmad[2], Nurfadilah[3]

60200117042@uin-alauddin.ac.id[1], 60200117056@uin-alauddin.ac.id[2], 60200117041@uin-alauddin.ac.id[3]

***Abstract*—During the Covid-19 pandemic, many people began to conflict over the application of the PSBB (Pembatasan Sosial Berskala Besar), ranging from difficulties when they wanted to leave the region, difficulties in the economy, or other problems. To find out people's opinions or problems regarding the implementation of the PSBB or (large periodic social restrictions) during the covid-19 pandemic period, we will do clustering, where later we will retrieve data from social media such as Twitter and then group the data according to its similarity using the K-Means Algorithm. The clustering process aims to minimize the occurrence of objective functions set in the clustering process which are generally used to minimize variations within a cluster. Where later we will determine the centroid value in each cluster and then calculate the closest distance using the Euclidean formula, so that it will find the appropriate results.**

***Kata Kunci*— Clustering, PSBB, K-Means, Covid-19.**

I. PENDAHULUAN

Virus corona jenis baru atau Covid-19 yang dalam istilah kedokteran disebut sebagai 2019 Novel Coronavirus (2019-nCoV) telah menyerang masyarakat dunia saat ini. Dikutip dari *Center for Disease Control and Prevention*, *cdc.gov*, virus corona merupakan jenis virus yang diidentifikasi sebagai penyebab penyakit pada saluran pernapasan, yang pertama kali terdeteksi muncul di Kota Wuhan, Tiongkok.

Virus ini diketahui pertama kali muncul di pasar hewan dan makanan laut di Kota Wuhan, Provinsi Hubei pada akhir Desember 2019. Setelah diketahui adanya virus ini, kasus masyarakat yang terkena virus ini terus melonjak hingga ke berbagai negara di dunia termasuk Indonesia [1].

Kasus positif corona di Indonesia pertama kali diumumkan awal bulan Maret 2020. Sejak saat itu, jumlah masyarakat yang positif corona terus bertambah setiap harinya, dan bahkan ada yang kehilangan nyawa. Karena banyaknya kasus positif tersebut, pemerintah kemudian melakukan berbagai cara atau aturan untuk menekan peningkatan kasus tersebut.

Salah satu cara yang diterapkan pemerintah untuk mencegah persebaran corona yakni dengan memberlakukan Pembatasan Sosial Berskala Besar (PSBB). PSBB diterapkan pertama kali di DKI Jakarta, kemudian disusul oleh daerah lain yang tingkat penyebaran atau kasusnya tinggi.

Semenjak diterapkannya PSBB, banyak memuai komentar-komentar atau opini mengenai kebijakan ini. Salah satu tempat bagi warga menyerukan pendapatnya yaitu melalui social media. Dengan melihat dan mengumpulkan opini netizen tersebut, kita dapat melakukan proses text mining dengan mengelompokkannya untuk mengetahui bagaimana pendapat mereka tentang PSBB. Untuk itu dilakukanlah clustering.

Clustering adalah proses pengelompokan titik-titik data kedalam dua kelompok atau lebih sehingga titik-titik data yang termasuk didalam kelompok yang sama lebih mirip satu sama lain daripada didalam kelompok yang berbeda, hanya berdasarkan informasi yang tersedia dengan poin data . Untuk melakukan pengelompokan data atau clustering, kami memilih algoritma k-means.

Algoritma K-Means merupakan salah satu dari algoritma yang banyak digunakan dalam pengelompokan karena kesederhanaan dan efisiensi dan diakui sebagai salah satu dari 10 algoritma data mining teratas oleh IEEE [2].

II. URAIAN PENELITIAN

A. *Text Mining*

Text mining adalah salah satu penambangan informasi yang berguna dari data – data yang berupa tulisan, dokumen atau text dalam bentuk klasifikasi maupun clustering. Text mining masih merupakan bagian dari data mining dimana akan memproses data – data atau text – text serta dokumen – dokumen yang bisa jadi dalam jumlah sangat besar. Untuk memproses data yang sangat besar tentulah akan memakan sumber daya yang tidak sedikit kaitanya dengan pengolahan data tersebut. Disinilah diperukanya sebuah pemrosesan awal atau preprocessing data text tersebut sebelum data tersebut di lakukan proses text mining sesuai algoritma yang akan diterapkan. Dengan *text mining* maka kita akan melakukan proses mencari atau penggalian informasi yang berguna dari data tekstual

Untuk dapat melakukan penambangan informasi atau text mining maka perlu dilakukan beberapa tahapan yang harus dilakukan untuk mengolah sumber data baik yang terstruktur, terstuktur sebagian dan yang tidak terstruktur dari beberapa sumber maka data-data tersebut perlu dilakukan proses awal atau di sebut sebagai preprocessing text yang bermaksud mengolah data awal yang masih bermacam–macam untuk dijadikan sebuah data teratur yang dapat dikenai atau diterapkan beberapa metode text mining yang ada [3].

B. *TF-IDF*

Metode *TF-IDF* merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada information retrieval. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat.

Metode *Term Frequency Inverse Document Frequency (TFIDF)* adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. TFIDF ini adalah sebuah ukuran

statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen atau dalam sekelompok kata. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Bobot kata semakin besar jika sering muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen [4].

Pada algoritma TFIDF digunakan rumus untuk menghitung bobot (W) masing masing dokumen terhadap kata kunci dengan rumus yaitu :

$$W_{dt} = tf_{dt} * IDF_t \qquad (1)$$

Di mana :
W_{dt} = bobot dokumen ke-d terhadap kata ke-t
tf_{dt} = banyaknya kata yang dicari pada sebuah dokumen
IDF_t = Inversed Document Frequency (log (N/df))
N = total dokumen
Df = banyak dokumen yang mengandung kata yang dicari.

Nilai IDF dihitung dengan :

$$IDF_j = \log\left(\frac{D}{df_i}\right) \qquad (2)$$

Di mana:
D = jumlah dokumen
Df = jumlah dokumen yang menganddung term (tj)

C. *Clustering*

Clustering merupakan contoh dari klasifikasi tanpa arahan (*unsupervised*). Klasifikasi merujuk kepada prosedur yang menetapkan objek data set kelas. Unsupervised berarti bahwa pengelompokkan tidak tergantung pada standar kelas dan pelatihan atau training.

Menurut Deka, *Clustering* merupakan salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau objek ke dalam *cluster* atau grup sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. *Clustering* melakukan pengelompokkan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode *unsupervised learning*. Menurut Oyelade, *clustering* dapat dibagi menjadi dua, yaitu *hierarchical clustering* dan *non-hierarchical clustering*.

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*. *Dendogram* biasanya digunakan untuk membantu memperjelas proses hierarki tersebut.

Berbeda dengan metode *hierarchical clustering*, metode *non-hierarchical clustering* justru dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan *K-Means Clustering* [5].

D. Algoritma *K-Means*

Algoritma *K-Means* merupakan salah satu algoritma dalam fungsi c*lustering* atau pengelompokan. *Clustering* mengacu pada pengelompokkan atas data, observasi atau kasus berdasarkan kemiripan objek yang diteliti. Sebuah *cluster* adalah suatu kumpulan data yang mirip dengan lainnya atau ketidakmiripan data pada kelompok lain [6]. *Clustering* didefinisikan dengan membagi objek data dalam bentuk, entitas, contoh, ketaatan, unit ke dalam beberapa jumlah kelompok (grup, bagian atau kategori).

Proses *clustering* bertujuan untuk meminimalkan terjadinya *objective function* yang diset dalam proses *clustering* yang pada umumnya digunakan untuk meminimalisasikan variasi dalam suatu *cluster* dan memaksimalkan variasi antar *cluster* atau dengan kata lain data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam kelompok lain.

Gambar 1. Flowchart K-Means Clustering.

Proses *clustering* dengan algoritma K-Means adalah sebagai berikut:

1. Tentukan banyaknya *cluster* yang diinginkan
2. Alokasikan data sesuai dengan jumlah *cluster* yang telah ditentukan
3. Tentukan nilai *centroid* pada tiap-tiap *cluster*
4. Hitung jarak terdekat dengan menggunakan rumus Euclidean
5. Tampilkan hasil berdasarkan jarak terendah dari hasil perhitungan step 4
6. Jika belum didapatkan hasil yang sesuai, iterasi kembali dilanjutkan dengan menggunakan step 3. Iterasi akan dihentikan jika hasil clustering sudah sama dengan iterasi sebelumnya [7].

Untuk menentukan nilai *centroid* tentukan bedasarkan nilai range yang berada pada sumber data yang ada dengan melakukan pemilihan sesuai dengan nilai *centroid* yang dipilih.

Untuk menentukan jarak digunakan rumus Euclidean sebagai berikut:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \qquad (3)$$

Di mana :

dist = Jarak Obyek
pk = Koordinat dari Obyek p
qk = Koordinat dari Obyek q
k = Urutan dari koordinat

III. PERANCANGAN SISTEM

A. Data

Data yang digunakan merupakan data yang diambil dari tweets netizen di social media twitter dari bulan April 2020 sampai bulan Mei 2020. Dengan menggunakan pemrograman bahasa *python* untuk men-*crawling* data pencarian kata yang berkaitan dengan "PSBB COVID".

Adapun jumlah data *crawling* yang diambil sebanyak 242 tweets. Contoh data tweets dapat dilihat pada Tabel 1.

TABEL I
CONTOH DATA TWEETS YANG DIDAPATKAN DARI HASIL CRAWLING

Nomor Tweets	Text Tweets
1	Di luar urusan yang birokratis ini, pemerintah daerah tetap mengupayakan berbagai cara untuk mencegah penyebaran Covid-19, walau tak diberi lampu hijau menerapkan PSBB #covid19 #COVID19indonesiahttps://www.bbc.com/indonesia/indonesia-52282767 …
2	Masyarakat Belum Paham Bahaya COVID-19, Anies Pastikan PSBB DKI Jakarta Diperpanjang! #CoronaVirus #AniesBaswedan

B. Text Preprocessing

Text Preprocessing adalah tahapan awal dari Text Mining untuk melakukan proses analisis terhadap suatu text dokumen. Proses Preprocessing dilakukan menggunakan R Studio dengan memanfaat beberapa library dan fungsi [8]. Adapun tahap preprocessing terdiri dari tahapan sebagai berikut :

1. *Case Folding*. Pada tahap ini text tweets akan diproses dengan merubah semua karakter huruf besar menjadi huruf kecil, selain itu juga menghilangkan beberapa karakter yang dianggap tidak valid seperti angka, tanda baca dan simbol. Proses *case folding*, perubahan text tweets menjadi huruf kecil dilakukan menggunakan library(tm) dengan fungsi 'dok_casefolding <- tm_map(corpusdok, content_transformer(tolower))'. Contoh hasil case folding lihat tabel II.

TABEL II
CONTOH HASIL PROSES CASE FOLDING

Nomor Tweets	Hasil Case Folding Text Tweets
1	di luar urusan yang birokratis ini, pemerintah daerah tetap mengupayakan berbagai cara untuk mencegah penyebaran covid-19, walau tak diberi lampu hijau menerapkan psbb #covid19 #covid19indonesiahttps://www.bbc.com/indonesia/indonesia-52282767â â€¦
2	masyarakat belum paham bahaya covid-19, anies pastikan psbb dki jakarta diperpanjang! #coronavirus #aniesbaswedan

2. *Text Cleaning*. Pada tahap ini akan dilakukan proses membersihkan text tweets, seperti menghapus *URL (Uniform Resource Locator)*, mention, hashtag, tanda baca, angka, simbol dan *slang word*. Proses *Text Cleaning* menggunakan library(tm) dengan fungsi tm_map.

TABEL III
CONTOH HASIL PROSES TEXT CLEANING

Nomor Tweets	Text Tweets
1	di luar urusan yang birokratis ini pemerintah daerah tetap mengupayakan berbagai cara untuk mencegah penyebaran covid walau tak diberi lampu hijau menerapkan psbb
2	masyarakat belum paham bahaya covid anies pastikan psbb dki jakarta diperpanjang

3. *Stemming*. Pada tahapan ini adalah proses merubah semua text tweets yang memiliki kata imbuhan menjadi kata dasar. Proses *stemming* menggunakan library (katadasaR) untuk bahasa Indonesia. [https://github.com/nurandi/katadasaR].

TABEL III
CONTOH HASIL PROSES STEMMING

Nomor Tweets	Hasil Stemming Text Tweets
1	di luar urus yang birokratis ini pemerintah daerah tetap upaya bagai cara untuk cegah

	sebar covid walau tak beri lampu hijau terap psbb
2	masyarakat belum paham bahaya covid anies pasti psbb dki jakarta panjang

4. *Stopword*. Pada tahapan ini text tweets akan diseleksi dengan menghilangkan kata-kata yang tidak memiliki nilai bobot atau makna yang disesuaikan dengan kamus stopword. Proses *stopword* dilakukan menggunakan library(tm) dengan fungsi "dok_stopword <- tm_map(dok_stemming, removeWords, cStopwordID)"

TABEL V
CONTOH HASIL PROSES STOPWORD

Nomor Tweets	Hasil Stopword Text Tweets
1	birokratis cegah sebar covid lampu terap psbb
2	bahaya covid anies psbb dki

5. *Tokenizing*. Pada tahap ini text tweets akan diproses dengan merubah kalimat pada text tweets menjadi potongan kata. Proses *tokenizing* dengan library(tm) menggunakan fungsi 'tdm = DocumentTermMatrix (data)'.

TABEL VI
CONTOH HASIL PROSES TOKENIZING

Nomor Tweets	Hasil Tokenizing Text Tweets
1	birokratis cegah covid lampu terap psbb
2	bahaya covid anies psbb dki

C. Text Representation

Text Representation merupakan tahap proses menghitung jumlah frekuensi term/kata pada data tweets dan merubah data tweets menjadi sebuah matriks yang memuat kolom jumlah term pada dokumen menggunakan library(tm) dengan fungsi dtm <- TermDocumentMatrix(data) dan tdm <- DocumentTermMatrix(data).

Berdasarkan hasil percobaan didapatkan jumlah frekuensi keseluruhan term dan jumlah term pada tiap dokumen. Lebih jelasnya lihat tabel VI dan VII.

TABEL VII
CONTOH HASIL DF (DOCUMENT FREQUENCY)

Term	Jumlah Term
covid	237
psbb	223
sebar	48

Gambar 2. Hasil wordcloud dari *document frequency*

TABEL VII
CONTOH HASIL TF (TERM FREQUENCY)

Nomor Tweets	Term				
	birokritas	cegah	covid	lampu	psbb
1	1	1	1	1	1
2	0	0	1	0	0

D. Term Weighting

Setelah melakukan text preprocessing dan representation akan menghasilkan term atau kata yang selanjutnya akan diproses term weighting atau menghitung nilai TF-IDF Perhitungan bobot tiap term dicari pada setiap dokumen yang bertujuan agar mampu mengetahui kesamaan atau kemiripan terhadap suatu term atau kata di dalam dokumen [7].

Dalam melakukan proses term weighting menggunakan metode pembobotan tf-idf yang kemudian dinormalisasikan. Berdasarkan hasil pengujian menggunakan library(tm) dengan fungsi "tdm.tfidf <- weightTfIdf(tdm)" didapatkan nilai term seperti tabel VII berikut.

TABEL VIII
CONTOH HASIL PROSES TF-IDF

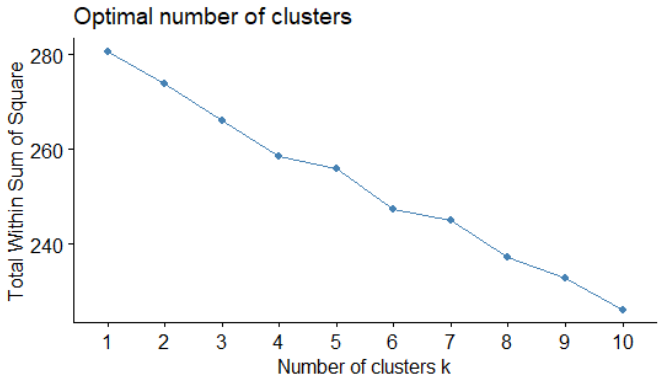
Nomor Tweets	Term				
	birokritas	cegah	covid	lampu	psbb
1	1.1312	0.3815	0.0159	1.1312	0.0292
2	0.0000	0.0000	0.0223	0.0000	0.0000

E. Penentuan Jumlah Kalster Terbaik

Untuk melakukan pencarian klaster terbaik menggunakan metode *elbow* untuk menduga nilai total wss (whitin sum square) sebagai penentu k optimalnya dengan nilai yang menunujukkan garis yang mengalami patahan membentuk elbow atau siku [9].

Berdasarkan hasil percobaan dengan menggunakan library(cluster) dengan fungsi "fviz_nbclust(data, kmeans, method = "silhouette")" diperoleh klaster optimal yang terbentuk seperti pada gambar 3.

Pada saat k=4, menunjukkan garis mengalami patahan yang membentuk elbow atau siku pada. Oleh karena itu dalam percobaan ini akan digunakan klaster dengan jumlah k=4.



Gambar 3. klaster terbaik menggunakan metode *elbow.*

IV. HASIL DAN ANALISIS

Dengan menggunakan metode *elbow* dalam menentukan jumlah klaster terbaik, maka didapatk jumlah klaster yang akan dicoba yakni dengan k=4. Selanjutnya dilakukan proses menghitung jarak *centroid* dari masing-masing dokumen menggunakan library(proxy) dengan fungsi "dist.matrix = dist(tfidf.matrix, method = "cosine")". Setelah itu dilakukan proses *clustering k-means* menggunakan library(cluster) dengan fungsi "cluster <- kmeans(data, 4, nstart = 25)".

Berdasarkan hasil *clustering k-means*, K=4 didapatkan jumlah tiap cluster. cluster 1 sebanyak 25, cluster 2 sebanyak 47, cluster 3 sebanyak 10 dan cluster 4 sebanyak 160. Lebih jelasnya lihat tabel IX.

TABEL IX
HASIL JUMLAH TWEETS SETIAP CLUSTER

Nomor Cluster	Jumlah Tweets
Cluster 1	25
Cluster 2	47
Cluster 3	10
Cluster 4	160

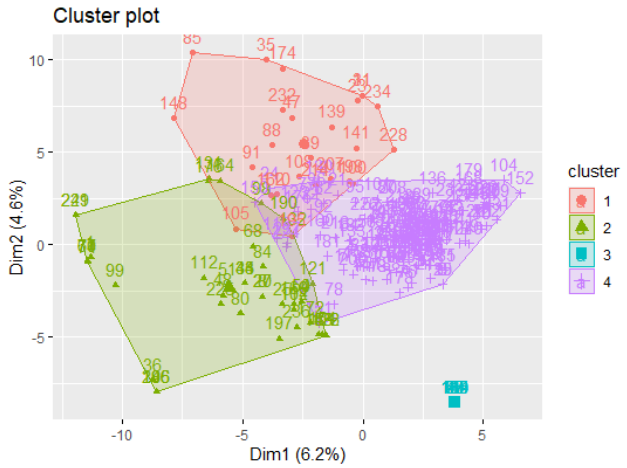
Untuk penentuan cluster pada setiap tweets pada dokumen didapatkan hasil seperti pada tabel VIII.

TABEL X
CONTOH HASIL PENENTUAN CLUSTER TEXT TWEETS

Nomor Tweets	Text Tweets	Nomor Cluster
1	Di luar urusan yang birokratis ini, pemerintah daerah tetap mengupayakan berbagai cara untuk mencegah penyebaran Covid-19, walau tak diberi lampu hijau menerapkan PSBB #covid19 #COVID19indonesiahttps://www.bbc.com/indonesia/indonesia-52282767 …	Cluster 2
2	Masyarakat Belum Paham Bahaya COVID-19, Anies Pastikan PSBB DKI Jakarta Diperpanjang! #CoronaVirus #AniesBaswedan	Cluster 4
3	Strategi Terbaru Perang COVID-19: Makassar Resmi Berstatus PSBB #merahputih #makassar #psbb https://merahputih.com/post/read/strategi-terbaru-perang-covid-19-makassar-resmi-berstatus-psbb …	Cluster 4
...	...	...
240	"Gini mas.. Warganya aja terlalu santuy gak ada panik panik nya pas ada isu covid ini nyebar di negara lain.. Sosoan indonesia kebal covid segala macem.. Udah ada pasien + aja masih ada yg bebal kalo dikasih tau.. Sekarang yg udah urgent aja warganya masih masa bodo psbb gak patuh"	Cluster 4
241	Berbagai ruas jalan di #jakarta tetep aja masih berdesak-desakan (dibaca:macet} dimana saat ini masih dlm waktu penerapan #psbb , untuk mencegah penyebaran pandemi #Covid_19 & sebenernya #psbbjakarta ini membuat Kota Jakarta menjadi lebih bersih udara.pic.twitter.com/Bm8itwj53Y	Cluster 2
242	Ayo ikuti ekspos terbuka hasil kajian "Dampak Ekonomi Covid-19". Rekomendasi PSBB dan implementasinya di Kota Makassar. Lawan Covid19pic.twitter.com/vZLsyk9UQY	Cluster 4

Hasil Analisis :

Pada gambar 4 terlihat nomor dokumen dikelompokkan dalam klaster tertentu. Dalam penentuan cluster didapatkan bahwa jumlah cluster paling banyak yakni cluster 4 dengan jumlah tweets 160 dan yang paling sedikit yakni cluster 3 dengan jumlah tweets 10.



Gambar 4. Hasil cluster plot text tweets.

Dari hasil rekapitulasi analisis text clustering didapatkan topik opini pembicaraan pada text tweets dikelompokkan sebagai berikut:

1. Cluster 1 : pada cluster 1 dikelompokkan berdasarkan kata yang paling sering muncul seperti mudik, kumpul, massa.
2. Cluster 2 : pada cluster 2 dikelompokkan berdasarkan kata yang paling sering muncul seperti protokol, patuh, bogor.
3. Cluster 3 : pada cluster 3 dikelompokkan berdasarkan kata yang paling sering muncul seperti bansos, bahaya, ramai.
4. Cluster 4: pada cluster 2 dikelompokkan berdasarkan kata yang paling sering muncul seperti sebar, cegah, pandemi, pasien.

V. KESIMPULAN

Dari pengujian atau penelitian yang telah dilakukan dapat disimpulkan bahwa :

1. Sebelum dilakukan clustering pada text tweets, terlebih dahulu harus dilakukan preprocessing diantaranya : *text cleaning, case folding, stemming, stopword dan tokenizing*.
2. Text tweets yang telah dipreprocessing kemudian akan dihitung nilai bobot TF-IDF.
3. Dengan menggunakan metode elbow untuk menduga nilai total wss (whitin sum square) sebagai penentu k optimalnya dengan nilai yang menunujukkan garis yang mengalami patahan membentuk elbow atau siku.
4. Hasil clustering k-means didapatkan dengan jumlah 4 cluster atau kelompok text tweets.
5. Cluster dengan jumlah tweets yang mendominasi dikelompokkan berdasarkan kata yang paling sering muncul seperti sebar, cegah, pandemi dan pasien

UCAPAN TERIMA KASIH

Kami menyampaikan rasa terima kasih yang sedalam-dalamnya kepada teman se-tim yang terlibat dalam perencanaan, penyusunan ide-ide hingga penyelesaian paper ini.

Kami juga menyampaikan terima kasih dan penghargaan yang setinggi-tingginya kepada Dosen Text Mining yang telah membimbing dan megarahkan kami untuk menyelesaikan laporan ini.

Selanjutnya kepada semua sumber referensi yang tertera dibawah ini, kami ucapkan terima kasih sebanyak-banyaknya atas ilmunya yang sangat bermanfaat.

REFERENSI

[1] Ariyanto. (2020, March 03). Asal Mula dan Penyebaran Virus Corona dari Wuhan ke Seluruh Dunia. Retrieved June 14, 2020, from https://bappeda.ntbprov.go.id/asal-mula-dan-penyebaran-virus-corona-dari-wuhan-ke-seluruh-dunia/

[2] X. Wu, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37

[3] Aris Tri Jaka H, ”Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining,” Jurnal Informatika UPGRIS Vol. 1, 2015.

[4] Putra, Agung Auliaguntary Arif. 2016. Implementasi Text Summarization Menggunakan Metode Vector Space Model pada Artikel Berita Bahasa Indonesia. Skripsi. Jurusan Teknik Informatika. Fakultas Teknik dan Ilmu Komputer. Universitas Komputer Indonesia

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

[6] Larose, D. T. (2005). An introduction todata mining. Traduction et adaptation deThierry Vallaud

[7] Yudi Agusta, "K-Means – Penerapan, Permasalahan," Jurnal Sistem dan Informatika, vol. 3, pp. 47-60, Februari 2007.

[8] Syarifah, L. (2019). Text Mining Untuk Pengklasifikasian Komentar Masyarakat Dalam Media Center Surabaya Dengan Metode Naïve Bayes Classifier (Doctoral Dissertation, Universitas Airlangga).

[9] Husain, A., A. (2018, February 2). Sign In. Retrieved June 25, 2020, from https://rpubs.com/ahmadhusain/355692