ANALISIS *CLUSTERING* TEKS TANGGAPAN MASYARAKAT DI TWITTER TERHADAP PEMBATASAN SOSIAL BERSKALA BESAR MENGGUNAKAN ALGORITMA K-MEANS

Muhammad Nur Akbar[1)], Darmatasia[2)], Mustikasari[3)], M. Syahwal[4)]
[1,2,3,4]Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Alauddin Makassar
[1,2,3,4]Jl. H. M. Yasin Limpo No. 36 Samata, Kab. Gowa, Sulawesi Selatan, Indonesia

E-mail: muhnurakbar@uin-alauddin.ac.id[1)], darmatasia@uin-alauddin.ac.id[2)],
mustikasari@uin-alauddin.ac.id[3)], 60200117042@uin-alauddin.ac.id[4)]

Abstrak – Virus corona (COVID-19) ditetapkan sebagai pandemi oleh WHO (*World Health Organization* atau Badan Kesehatan Dunia) karena penyebarannya yang terus meningkat dan telah mencapai sebagian besar negara di dunia, termasuk Indonesia. Setiap negara dituntut dapat lebih agresif dalam mengambil tindakan pencegahan dan perawatan. Pemerintah Indonesia sendiri mengeluarkan kebijakan berupa wajib masker, jam malam, serta PSBB (Pembatasan Sosial Berskala Besar) guna menekan laju menyebaran COVID-19. Namun kebijakan tersebut menuai tanggapan pro dan kontra dari masyarakat khususnya melalui media sosial, di satu sisi PSBB dianggap mampu menekan laju penyebaran COVID-19 namun di sisi lain PSBB dianggap akan memperburuk kondisi perekonomian masyarakat, khususnya golongan menengah bawah. Penelitian ini bertujuan untuk mengelompokkan tanggapan masyarakat mengenai PSBB di twitter ke dalam beberapa *cluster*, tanggapan yang berada dalam satu *cluster* yang sama dianggap memiliki topik atau karakteristik pembahasan yang serupa dan sebaliknya, sehingga dapat memberi *insight* tambahan pada pihak pemerintah dalam mengevaluasi kebijakannya. Algoritma K-Means digunakan untuk mengelompokkan tanggapan yang memiliki kesamaan karakteristik sebab terbukti memiliki tingkat akurasi yang tinggi dengan waktu eksekusi yang relatif cepat karena bersifat linear. Penelitian ini menghasilkan 4 *cluster* berbeda dengan mengunakan metode Elbow dalam penentuan jumlah K pada algoritma K-Means dan nilai SSE (*Sum of Square Error*) sebagai parameter evaluasinya.

Kata Kunci: COVID-19, PSBB, Clustering, Twitter, K-Means.

Abstract *– The coronavirus (COVID-19) is designated as a pandemic by the WHO (World Health Organization) because of its ever-increasing spread and has reached most countries in the world, including Indonesia. Each country is required to be more aggressive in taking preventive and treatment measures. The Indonesian government issued policies in the form of mandatory masks, curfews, and PSBB (Large-Scale Social Restrictions) to suppress the spread of COVID-19. However, this policy has drawn pro and contra responses from the public, especially through social media. On the one hand, PSBB is considered capable of suppressing the spread of COVID-19, but on the other hand, PSBB is considered to worsen the economic condition of the community, especially the lower middle class. This study aims to classify public responses regarding PSBB on Twitter into several clusters, responses that are in the same cluster are considered to have similar topics or characteristics of discussion and vice versa, so that they can provide additional insight to the government in evaluating its policies. The K-Means algorithm is used to group responses that have similar characteristics because it is proven to have a high level of accuracy with a relatively fast execution time because it is linear. This study resulted in 4 different clusters using the Elbow method in determining the number of K in the K-Means algorithm and the SSE (Sum of Square Error) value as the evaluation parameter.*

Keywords: *COVID-19, PSBB, Clustering, Twitter, K-Means.*

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 (Ariyanto, 2020).

Virus corona (COVID-19) kemudian ditetapkan sebagai pandemi oleh WHO (*World Health Organization* atau Badan Kesehatan Dunia) karena penyebarannya yang terus meningkat dan telah

mencapai sebagian besar negara di dunia termasuk Indonesia. COVID-19 ditetapkan sebagai pandemi agar negara-negara di dunia bisa lebih agresif dalam mengambil tindakan pencegahan dan perawatan.

Kasus positif corona di Indonesia sendiri pertama kali diumumkan awal bulan Maret 2020. Sejak saat itu, jumlah masyarakat yang positif corona terus meningkat setiap harinya, dan bahkan ada yang kehilangan nyawa. Tercatat setidaknya 210.910 orang terkonfirmasi positif dan sebanyak 8.544 orang meninggal akibat COVID-19 hingga tanggal 11 September 2020 (Gugus Tugas Percepatan Penanganan COVID-19, 2020). Pemerintah Indonesia kemudian menanggapi hal tersebut dengan penanganan yang dilakukan melalui berbagai kebijakan dan aturan guna menekan laju menyebaran COVID-19, yakni dengan memberlakukan Pembatasan Sosial Berskala Besar (PSBB), aturan wajib masker, serta aturan jam malam. PSBB diterapkan pertama kali di DKI Jakarta, kemudian disusul oleh daerah lain yang tingkat penyebaran atau kasusnya tinggi. Kebijakan-kebijakan tersebut diharapkan dapat memastikan masyarakat untuk tetap menerapkan protokol kesehatan pada masa adaptasi kebiasaan baru atau disebut *new normal*.

Namun kebijakan-kebijakan yang ada dianggap belum berjalan efektif yang tercermin dari tanggapan pro dan kontra masyarakat khususnya melalui media sosial, di satu sisi PSBB dianggap mampu menekan laju penyebaran COVID-19 namun di sisi lain PSBB dianggap akan memperburuk kondisi perekonomian masyarakat, khususnya golongan menengah bawah. Hal tersebut juga diperparah dengan munculnya fenomena matinya kepakaran (*the death of expertise*) yang berkembang di masyarakat. Dalam buku Matinya Kepakaran, Tom Nichols menyatakan bahwa fenomena ini dapat diamati dari sikap masyarakat yang memiliki kecenderungan tidak percaya terhadap penyataan pakar yang kemudian tercermin dari sikap bebal dalam mengikuti kebijakan permerintah (T. Nichols, 2018).

Pemerintah Indonesia perlu memperhatikan pelaksanaan kebijakan mengingat pelaksanaan prosedur yang tidak sesuai justru dapat memicu meningkatnya ketidakpercayaan masyarakat pada pemerintah.

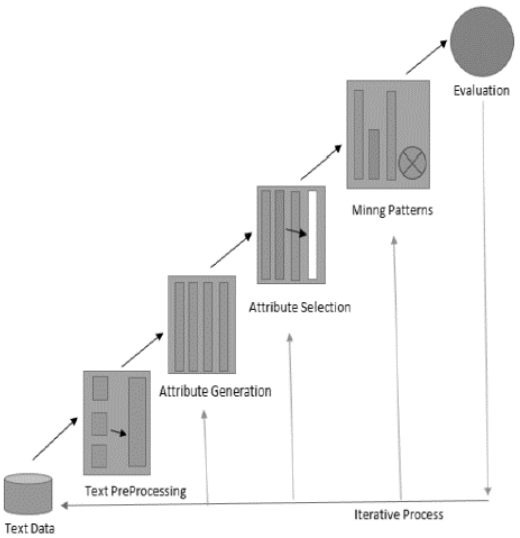
Di satu sisi masyarakat seringkali mengutarakan tanggapannya terkait suatu kebijakan melalui sosial media, tidak terkecuali pada Twitter. Sehingga pemerintah dapat memanfaatkan sosial media dalam memonitor dan mengevaluasi kebijakan yang sedang berjalan dengan melakukan analisa terhadap tanggapan masnyarakat di sosial media (Dirjan IKP Kementerian Komunikasi dan Informatika, 2018). Di sisi lain *text mining* memungkinkan adanya pengolahan data teks otomatis dalam jumlah besar untuk menemukan *insight* atau wawasan baru dari sekumpulan tanggapan masyarakat terhadap kebijakan yang berlaku. Hasil tersebut diharapkan dapat dijadikan bahan evaluasi dan acuan dalam penyusunan kebijakan berikutnya.

Clustering merupakan salah satu metode analisis dalam *text mining* yang digunakan untuk mengelompokkan data teks ke dalam dua kelompok atau lebih sehingga data-data yang termasuk di dalam kelompok yang sama akan memiliki kemiripan karakteristik satu sama lain daripada kelompok yang berbeda. Dalam penelitian ini data teks yang dimaksud berupa data tanggapan atau komentar masyarakat terkait kebijakan PSBB pada sosial media Twitter. Untuk melakukan analisis *custering* digunakan Algoritma K-Means. Algoritma K-Menas merupakan salah satu dari algoritma yang banyak digunakan dalam pengelompokan data karena sederahana dan tingkat efisiensi yang baik serta diakui sebagai salah satu dari sepuluh algoritma data mining teratas oleh IEEE (Wu, 2008).

METODOLOGI PENELITIAN

1. Tahapan Penelitian

Analisis *clustering* teks pada penelitian ini terdiri dari beberapa tahapan mengikuti tahapan pada proses *text mining*, yaitu *text preprocessing*, *attribute generation*, *attribute selection*, *mining patterns*, dan *evaluation*.



Gambar 1 Tahapan *Text Mining* (Kumar, 2013)

2. Data

Data yang digunakan merupakan data yang diambil dari tweets masyarakat di media sosial twitter dari bulan April 2020 sampai bulan Mei 2020. Dengan menggunakan pemrograman bahasa python dalam melakukan *crawling* data dengan 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 1 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 diberilampu 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

3. Text Mining

Text Mining didefinisikan sebagai ekstraksi *non-trivial* dari informasi tersembunyi, yang sebelumnya tidak diketahui, dan berpotensi berguna dari (sejumlah besar) data tekstual (Waegel, 2006). *Text Mining* merupakan bidang baru yang mencoba mengekstrak informasi yang bermakna dari teks bahasa alami. Dapat didefinisikan pula sebagai proses menganalisis teks untuk mengekstrak informasi yang berguna untuk tujuan tertentu. Dibandingkan dengan jenis data yang disimpan dalam database, teks meiliki karakteristik tidak terstruktur, ambigu, dan sulit untuk diproses. Namun demikian, pada masa modern saat ini, teks adalah cara paling komunal untuk pertukaran informasi secara formal. *Text mining* biasanya berkaitan dengan teks berupa komunikasi informasi aktual atau pendapat, dan rangsangan untuk mencoba mengekstrak informasi dari teks tersebut secara otomatis.

Text Mining serupa dengan *data mining*, kecuali bahwa *tools data mining* (Navathe, 2000) hanya dirancang untuk menangani data terstruktur dari database, sedangkan *tools text mining* juga dapat bekerja dengan kumpulan data tidak terstruktur atau semi-terstruktur seperti email, dokumen teks, dan file HTML, dll. Sehingga *text mining* adalah dianggap solusi yang jauh lebih baik.

4. Text Preprocessing

Text preprocessing merupakan salah satu komponen dalam *text mining*. *Text preprocessing* dilakukan untuk mengubah data tekstual yang tidak terstruktur ke dalam data yang terstruktur dan disimpan ke dalam basis data (Langgeni dkk., 2010). Tujuan dari preprocessing yakni menghasilkan sebuah *set term index* yang bisa mewakili dokumen. Komponen dari text *preprocessing* dibagi menjadi beberapa bagian, yaitu:

a. Tokenisasi

Tokenisasi adalah tugas memisahkan deretan kata di dalam kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal atau *termmed word*. Pada saat bersamaan, tokenisasi juga membuang beberapa karakter tertentu yang dianggap sebagai tanda baca.

b. 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.

c. Penghilangan *Stopword*

Stopword didefinisikan sebagai term yang tidak berhubungan (*irrelevant*) dengan subyek utama dari database meskipun kata tersebut sering kali hadir di dalam dokumen. Berikut ini adalah contoh *stopwords* dalam bahasa Indonesia: yang, juga, dari, dia, kami, kamu, aku, saya, ini, itu, atau, dan, tersebut, pada, dengan, adalah, yaitu, ke, tak, tidak, di, pada, jika, maka, ada, pun, lain, saja, hanya, namun, seperti, kemudian, dll.

d. *Stemming*

Kata-kata yang muncul di dalam dokumen sering mempunyai banyak varian morfologik. Karena itu, setiap kata yang bukan stop-words direduksi ke bentuk *stemmed word* (term) yang cocok. Kata

tersebut distem untuk mendapatkan bentuk akarnya dengan menghilangkan awalan atau akhiran. Dengan cara ini, diperoleh kelompok kata yang mempunyai makna serupa tetapi berbeda wujud sintaktis satu dengan lainnya.

Kelompok tersebut dapat direpresentasikan oleh satu kata tertentu. Sebagai contoh, kata menyebutkan, tersebut, disebut dapat dikatakan serupa atau satu kelompok dan dapat diwakili oleh satu kata umum sebut.

5. Term Weighting

Term Weighting adalah suatu pembobotan kata dalam suatu dokumen yang sering digunakan dalam proses text mining (Asian, 2007).

a. TF-IDF

Term Frequency (TF) merupakan frekuensi kemunculan term pada dokumen. TF suatu dokumen dengan dokumen yang lain akan berbeda, bergantung pada tingkat kepentingan sebuah term dalam dokumen. *Inverse Document Frequency* (IDF) merupakan sebuah perhitungan bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan atau dengan kata lain menghitung sebuah keunikan sebuah term dalam sebuah dokumen dibandingkan dengan dokumen yang lain. Semakin sedikit dokumen yang mengandung term yang dimaksud, maka nilai idf semakin besar. Jika setiap dokumen dalam koleksi mengandung term yang bersangkutan, maka nilai dari idf dari term tersebut adalah nol. Hal ini menunjukkan bahwa sebuah term yang muncul pada setiap dokumen dalam koleksi tidak berguna untuk membedakan dokumen berdasarkan topik tertentu. Nilai IDF sebuah term t dirumuskan dalam persamaan berikut:

$$IDF(t) = \log (N/df(t)) \qquad (1)$$

N adalah jumlah dokumen dan df(t) adalah jumlah dokumen yang mengandung term yang bersangkutan.

Dengan menggunakan tf-idf maka dapat diketahui deskripsi terbaik dari dokumen adalah term yang banyak muncul dalam dokumen tersebut dan sangat sedikt kemunculannya pada dokumen yang lain. Bobot terendah akan diberikan pada term yang muncul sangat jarang pada beberapa dokumen (*low-frequency documents*) dan term yang muncul pada hampir atau seluruh dokumen (*high-frequency documents*). Penelitian belakangan ini (Salton, 1989) telah mengkombinasikan TF dan IDF untuk menghitung bobot term dan menunjukkan bahwa gabungan keduanya menghasilkan performansi yang lebih baik. Kombinasi bobot dari sebuah term t pada text d didefinisikan sebagai berikut :

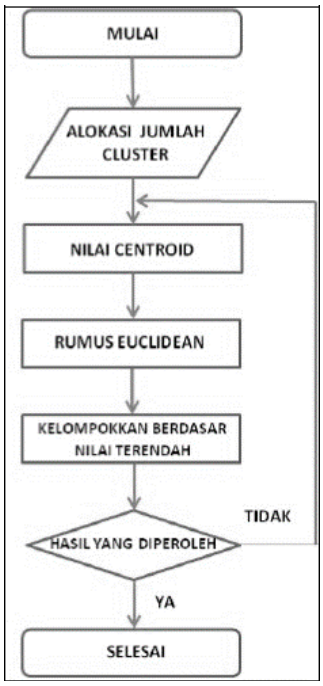
$$TFIDF(d.t) = TF(d.t).IDF(t) \qquad (2)$$

Faktor TF dan IDF dapat berkontribusi untuk memperbaiki nilai akurasi, *recall* dan *precision*.

6. K-Means Clustering

Algoritma K-Means merupakan salah satu algoritma dalam fungsi *clustering* 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 (Larose, 2005). *Clustering* didefinisikan dengan membagi objek data dalam bentuk, entitas, contoh, ketaatan, unit ke dalam beberapa jumlah kelompok (grup, bagian atau kategori).

Proses *clustering* bertujuan 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 memiliki karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan data yang memiliki karakteristik berbeda akan dikelompokkan ke dalam kelompok lain.



Gambar 2 *Flowchart* K-Means Clustering

Proses *clustering* dengan algoritma K-Means adalahsebagai berikut:

1. Tentukan banyaknya *cluster* yang diinginkan
2. Alokasikan data sesuai dengan jumlah *cluster* yang telahditentukan
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 (Syarifah, 2019).

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

Dalam menentukan jarak antar dua titik digunakan rumus Euclidean:

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

7. Metode Elbow

Metode Elbow adalah metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik (Merliana, 2015).

Berikut ini tahapan algoritma metode Elbow dalam menentukan nilai k pada K-Means:

1. Menginisialisasi awal nilai k ;
2. Menaikan nilai k ;
3. Menghitung hasil *sum of square error* dari tiap nilai k ;
4. Analisis hasil *sum of square error* dari nilai k yang mengalami penurunan secara drastis ;
5. Cari dan tetapkan nilai k yang berbentuk siku.

Pada metode Elbow nilai cluster terbaik yang akan diambil dari nilai *Sum of Square Error* (SSE) yang mengalami penurunan yang signifikan dan berbentuk siku. Untuk menghitung SSE menggunakan rumus:

$$SSE = \sum_{K=1}^{K}\sum_{x_i s S_k}\|X_i - C_k\|^2 \qquad (4)$$

dimana:

k = jumlah cluster
xi = data ke – *i*
Ck = *centroid cluster*

Sum of Square Error (SSE) merupakan rumus yang digunakan untuk mengukur perbedaan antara data yang diperoleh dengan model perkiraan yang telah dilakukan sebelumnya. SSE sering digunakan sebagai acuan penelitian terkait dalam menentukan optimal *cluster*.

HASIL DAN PEMBAHASAN

1. 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. Adapun tahap preprocessing terdiri dari tahapan sebagai berikut :

a. 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 2.

Tabel 2 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

b. 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 3 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

c. 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 4 Contoh Hasil Proses *Stemming*

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

d. 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 5 Contoh Hasil Proses *Stopword*

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

e. 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 6 Contoh Hasil Proses *Tokenizing*

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

2. Term 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 7 dan 8.

Tabel 7 Contoh Hasil DF (*document frequency*)

Term	Jumlah Term
covid	237
psbb	223
sebar	48

Gambar 3 Hasil *wordcloud* dari *document frequency*

Tabel 8 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

3. *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 (Yudi, 2007).

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 9 berikut.

Tabel 9 Contoh Hasil Proses TF-IDF

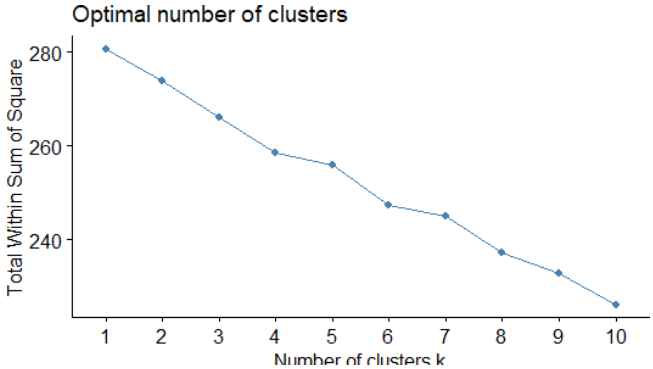
Nomor Tweets	Term				
	birokritas	cegah	covid	lampu	psbb
1	1.1312	0.3815	0.0159	1.1312	0.0292
2	0.0000	0.0000	0.0223	0.0000	0.0000

4. Penentuan Jumlah *Cluster* 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 (Husein, 2018).

Berdasarkan hasil percobaan dengan menggunakan library(cluster) dengan fungsi "fviz_nbclust(data, kmeans, method = "silhouette")" diperoleh klaster optimal yang terbentuk seperti pada gambar 4.

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 4 Penentuan jumlah cluster dengan metode Elbow

5. Hasil *Clustering*

Dengan menggunakan metode Elbow dalam menentukan jumlah klaster terbaik, maka didapatkan 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 10.

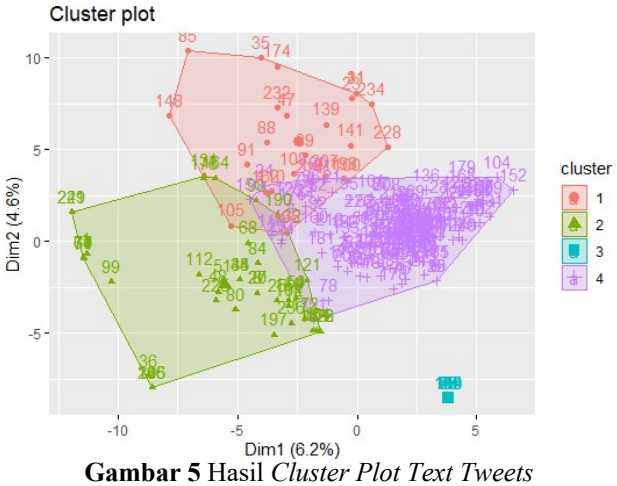
Tabel 10 Hasil Jumlah Tweets Setiap Cluster

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 11.

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 masihdlm waktu penerapan #psbb , untuk mencegah penyebaran pandemi #Covid_19 & sebenernya #psbbjakarta ini membuat Kota Jakarta menjadi lebih bersih udara.pic.twitter.com/Bm8itwj 53Y	Cluster 2
242	Ayo ikuti ekspos terbuka hasil kajian"Dampak Ekonomi Covid-19". Rekomendasi PSBB dan implementasinya di Kota Makassar. Lawan Covid19pic.twitter.com/vZLsyk9 UQ Y	Cluster 4

Pada gambar berikut 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 5 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 4 dikelompokkan berdasarkan kata yang paling sering muncul seperti sebar, cegah, pandemi, pasien.

KESIMPULAN

Dari pengujian 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*. Text tweets yang telah di*preprocessing* kemudian akan dihitung nilai bobot TF-IDF.
2. 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. Ditemukan nilai k terbaik = 4.

3. Hasil clustering k-means didapatkan dengan jumlah 4 cluster atau kelompok text tweets. Cluster dengan jumlah tweets yang mendominasi dikelompokkan berdasarkan kata yang paling sering muncul seperti sebar, cegah, pandemi dan pasien.
4. Penelitian ini dapat dikombinasikan dengan *sentiment analysis* pada pengembangan penelitian berikutnya.

DAFTAR PUSTAKA

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-coronadari-wuhan-ke-seluruh-dunia/

Gugus Tugas Percepatan Penanganan COVID-19. 2020, "Infografis COVID-19," *Infografis COVID-19*, 2020. Https://Covid19.go.id/ (diakses Sep. 11, 2020).

T. Nichols, Matinya Kepakaran. Jakarta: Gramedia Pustaka Utama, 2018.

Dirjen IKP Kementerian Komunikasi dan Informatika. 2018. "Memaksimalkan Penggunaan Sosial Media Dalam Lembaga Pemerintah". Jakarta, 10 Desember 2018.

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

Kumar, Lokesh., Bathia, Parul K. 2013. Text Mining : Concept, Process, and Applications. *Journal of Global Research in Computer Science Vol 4 No 3 March 2013*.

Waegel, Daniel. 2006. "The Development of Text Mining Tools and Algorithms". Ursinus College, 2006.

Navathe, Shamkant B., Ramez, Elmasri. 2000. "Data Warehousing and Data Mining" in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, 841-872, 2000.

Langgeni, Baizal & Firdaus. 2010. Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Yogyakarta : Seminar Nasional Informatika.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-wesley, Reading, Pennsylvania..

Larose, D. T. 2005. An introduction todata mining. Traduction et adaptation deThierry Vallaud

Asian, J. 2007. Effective Techniques for Indonesian Text Retrieval. PhD. Royal Melbourne Institute of Technology University.

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

N. P. E. Merliana, Ernawati dan A. J. Santoso. 2015. "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K- Means," UNISBANK , 2015.

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