IMPLEMENTASI *TEXT MINING* UNTUK *ADVERTISING* DENGAN MENGGUNAKAN METODE *K-MEANS CLUSTERING* PADA DATA *TWEET*S GOJEK INDONESIA

Azizah Nurfauziah Yusril, Inggrit Larasati, Qurrotul Aini
Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah
Jl. Ir H. Juanda No.95, Cempaka Putih, Kecamatan Ciputat, Kota Tangerang Selatan, Banten
Email: azizah.ny17@mhs.uinjkt.ac.id, inggrit.larasati17@mhs.uinjkt.ac.id, qurrotul.aini@uinjkt.ac.id

(Diterima: 10 Juli 2020, direvisi: 15 Agustus 2020, disetujui: 31 Agustus 2020)

ABSTRACT

In determining advertising, businesses use social media to find out the responses of their followers. Gojek Indonesia is one of the company in Indonesia who uses Twitter social media as a means to do advertising. The purpose of this research is to find out the type of tweet content that is mostly retweeted and favorite by Gojek Indonesian followers so that it can be used to do advertising to Twitter users. The collection of tweet data from Twitter is done by integrating Twitter API and R programming language using R Studio tools. The data analysis method uses text mining and for the clustering process uses K-means. The results of this study obtained a number of 2 cluster tweets. Based on the calculation of the average number of retweets in each cluster, it was found that the type of content with the most retweets was related to the quiz program and the introduction of Gojek Indonesia's products. Gojek Indonesia business people can use the retweet and favorite features as a means of advertising Twitter users.

Keywords: *advertising, gojek indonesia, text mining, k-means clustering, twitter*

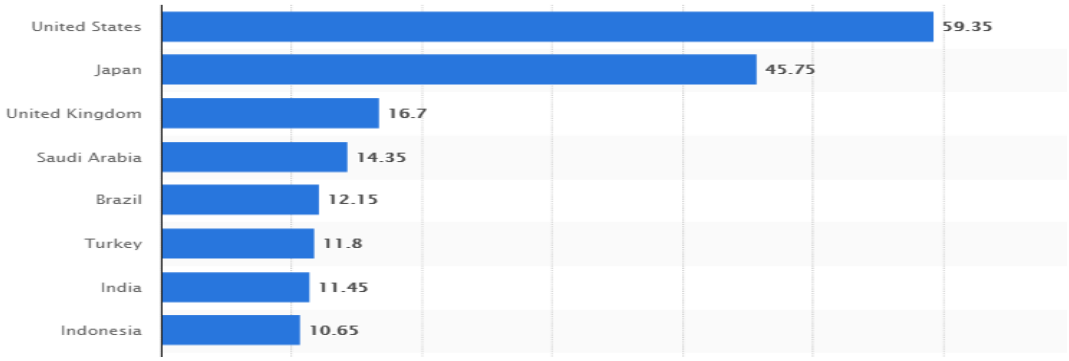
ABSTRAK

Dalam penentuan *advertising*, pelaku bisnis menggunakan media sosial untuk mengetahui respon dari para *followers*-nya. Gojek Indonesia merupakan salah satu pelaku bisnis di Indonesia yang menggunakan media sosial Twitter sebagai sarana untuk melakukan *advertising*. Tujuan dari penelitian ini yaitu untuk mengetahui jenis konten *tweets* yang banyak dilakukan *retweet* dan *favorite* oleh *followers* Gojek Indonesia sehingga dapat digunakan untuk melakukan *advertising* kepada pengguna Twitter. Pengumpulan data *tweets* dari Twitter dilakukan dengan mengintegrasikan Twitter API dan bahasa pemograman R menggunakan *tools* R Studio. Metode analisis data menggunakan *text mining* dan untuk klasterisasi menggunakan *K-means*. Hasil dari penelitian ini didapatkan sejumlah 2 klaster *tweets*. Berdasarkan perhitungan jumlah rata-rata *retweet* pada tiap klaster, didapatkan bahwa jenis konten dengan *retweet* terbanyak yaitu terkait program kuis dan perkenalan produk Gojek Indonesia. Pelaku bisnis Gojek Indonesia dapat menggunakan fitur *retweet* dan *favorite* sebagai sarana untuk melakukan a*dvertising* kepada pengguna Twitter.

Kata Kunci: *advertising*, gojek indonesia, *text mining*, *k-means clustering*, twitter

1 PENDAHULUAN

Sebagian besar pelaku bisnis menggunakan media sosial sebagai sarana untuk melakukan *advertising*. Pada gambar 1 yang diperoleh dari Statista, pengguna Twitter di Indonesia menempati urutan ke-8 di dunia dengan jumlah kurang lebih 10 juta pengguna. Twitter bisa digunakan untuk mendukung sosial media *networking* untuk subjek pemasaran, iklan, jurnalistik atau komunikasi subjek [1]. Sosial media analitik bisa memberikan informasi berharga bagi para pemimpin perusahaan untuk membuat keputusan strategis dan mengatasi masalah-masalah bisnis [2]. Gojek Indonesia merupakan salah satu pelaku bisnis di Indonesia yang bergerak di bidang jasa transportasi yang menggunakan media sosial Twitter sebagai sarana untuk melakukan *advertising.*



Gambar 1. Grafik Pengguna Aktif Twitter Per April 2020 [3]

Terhitung pada tanggal 9 April 2020, Gojek Indonesia memiliki jumlah *followers* Twitter sebanyak 946 ribu *followers*, dan jumlah *tweets* sebanyak 482 ribu. *Advertising* di Twitter bisa dilakukan dengan melakukan *tweet* promosi, akun promosi dan tren promosi. *Tweet* promosi adalah *tweet* biasa yang diiklankan oleh pengiklan dengan membayar penempatannya di Twitter. Twitter menentukan *tweet* promosi mana yang relevan bagi pengguna berdasarkan *tweet* pengguna, siapa saja yang di-*follow*, apa yang di-*retweet*, apa yang dicari, apa yang dilihat dan interaksi dengan *tweet* atau akun di Twitter.

Dengan fitur *retweet*, Gojek Indonesia bisa menerapkan *viral marketing*, strategi *marketing* yang dilakukan dengan menyebarkan infomasi produk atau layanan melalui strategi "*word of mouth*" atau komunikasi dari mulut ke mulut dengan memanfaatkan kecepatan penyebaran informasi. Dilihat dari aspek *cost*, strategi ini lebih baik karena tidak membutuhkan biaya dibanding menggunakan *tweet* promosi yang disediakan Twitter. Strategi ini dikatakan berhasil ketika pengguna terdorong untuk membagikan kembali informasi tersebut, misalnya dengan me-*retweet* kembali *tweet* yang dibuat oleh akun Gojek Indonesia.

Oleh karena itu penting mengetahui jenis konten *tweet* apa yang banyak dilakukan *retweet* oleh *followers* Gojek Indonesia. Untuk mengetahui konten iklan apa yang paling diminati oleh pengguna Twitter bisa dilakukan dengan cara *text mining* terhadap data *tweets* dengan menerapkan teknik *clustering*. Salah satu metode yang digunakan dalam *clustering* adalah metode *k-means*. Pengumpulan data *tweets* dari Twitter dapat dilakukan dengan mengintegrasikan Twitter API dan bahasa pemograman R dengan *tools* R Studio.

Text mining merupakan analisis teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan, keterkaitan dan kelas antar dokumen [9]. *Text mining* dapat didefinisikan secara luas sebagai suatu proses menggali informasi yang berasal dari sekumpulan dokumen dari waktu ke waktu menggunakan serangkaian alat analisis untuk mengidentifikasi dan mengeksplorasi pola data yang ada. Pada dasarnya *text mining* memiliki konsep pengolahan yang hampir sama dengan *data mining*, perbedaannya yaitu terdapat pada sumber data yang digunakan. Sumber data *text mining* berupa teks tidak terstruktur, sedangkan *data mining* menggunakan data terstruktur [10]. *Text mining* melingkupi sebuah proses ekstraksi informasi yang terpola yang berasal dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipan teks, atau bahkan SMS [11].

Tujuan dari penelitian ini adalah untuk mengetahui jenis konten *tweet* yang banyak dilakukan *retweet* dan *favorite* oleh *followers* Gojek Indonesia sehingga dapat digunakan sebagai sarana *advertising* kepada pengguna Twitter. Pengumpulan data *tweets* dari Twitter dilakukan dengan mengintegrasikan Twitter API dan bahasa pemograman R dengan *tools* R Studio, metode analisis *text mining* dengan menggunakan *k-means*.

2 TINJAUAN PUSTAKA

Pada artikel sejenis [4] dilakukan penerapan *text mining* untuk melakukan *clustering* dengan metode *k-means* pada data *tweets* Shopee Indonesia untuk mengetahui jenis konten *tweet* yang banyak dilakukan *retweet* oleh *followers* Shopee Indonesia. Hasil yang didapat adalah jenis konten pada klaster yang memiliki jumlah *retweet* yang tinggi diantaranya tentang kuis berhadiah (klaster 4,

klaster 20, dan klaster 28), ulang tahun Shopee Indonesia (klaster 11), dan hobi, kuis, dan *lifestyle* (klaster 7), sehingga Shopee Indonesia menggunakan jenis konten *tweet* tersebut untuk melakukan *advertising*. Pada artikel yang masih membahas *e-commerce* [5] aplikasi *text mining* ke klaster *tweet* dari akun Twitter @LazadaID dengan menggunakan algoritma pengelompokan *modified gustafson-kessel*. Hasil penelitian menunjukkan bahwa jumlah klaster optimal yang dibentuk berdasarkan indeks validasi partisi dan klasifikasi entropi klasifikasi adalah tiga klaster yang berisi penawaran barang elektronik, diskon, dan potongan harga. *Tweet* dengan *retweet* dan *favorite* terbanyak adalah *tweet* kuis hadiah. PT Lazada Indonesia dapat menggunakan *tweet* semacam ini untuk melakukan iklan di Twitter karena kuis hadiah disukai oleh pengikut akun Twitter @LazadaID. Pada artikel sejenis dengan objek *e-commerce* Lazada [6], pengumpulan data *tweets* dari Twitter dapat dilakukan dengan mengintegrasikan Twitter API dan RapidMiner. Metode analisis data menggunakan algoritma *classic naive bayes*. Hasil analisis menunjukkan hasil yang signifikan pada analisis sentimen dengan tingkat akurasi dari 98,29%. Selanjutnya pada [7] digunakan dataset dari 10 pantai yang ada di Indonesia sebanyak 500 *tweets*. Hasil akurasi dari klasifikasi menggunakan algoritma *support vector machine* sebesar 74,39%. Selanjutnya data opini dari kuesioner ditambahkan untuk mengelompokkan pantai berdasarkan ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar dan posisi pariwisata. Proses pengelompokan data ini menggunakan metode *k-means*. Kemudian artikel [8] *text clustering* digunakan untuk mengelompokkan pendapat menjadi beberapa kategori. Metode yang digunakan adalah metode *k-means* dan *Density Based Spatial Clustering of Applications with Noise* (DBSCAN). Berdasarkan nilai *silhouette coefficient*, metode DBSCAN lebih baik daripada *k-means* dalam mengelompokkan *tweet* yang ditujukan kepada layanan ekspedisi JNE, J&T, dan Pos Indonesia karena menghasilkan *silhouette coefficient* yang lebih tinggi.

Dari beberapa penelitian sebelumnya yang telah disebutkan diatas, *k-means* terbukti memiliki keuntungan yang lebih dibandingkan metode *clustering* yang lain karena tidak memerlukan iterasi yang banyak dalam proses *clustering*. Sehingga pada penelitian kami, *k-means* sangat cocok digunakan untuk mengetahui jenis konten *tweet* yang banyak dilakukan *retweet* dan *favorite* oleh *followers* Gojek Indonesia untuk dapat digunakan sebagai sarana *advertising*.

3 METODE PENELITIAN

Penelitian ini menggunakan metode penelitian kualitatif. Data yang digunakan yaitu data primer dan data sekunder. Data primer yang digunakan adalah dataset *tweet* dari Twitter. Sedangkan, data sekunder yang digunakan berupa buku, jurnal dan prosiding.

Tahap pertama di dalam penelitian ini adalah *authentication* untuk proses integrasi antara Twitter API dengan R. Oleh karena itu, peneliti harus mendapatkan API *keys* dan *tokens* dari Twitter terlebih dahulu dengan melakukan beberapa pengaturan di dalam *platform developer* Twitter yaitu *https://developer.Twitter.com/en/apps*. Pengaturan pertama yaitu membuat aplikasi baru dengan mengisi data-data pendukung. Setelah itu, membuat *token* yang diperoleh dari akun Twitter dan menyalin kode *consumer key (*API *key), consumer secret, access token* dan *access token secret* ke dalam aplikasi yang telah dibuat. Tahap kedua, melakukan *data acquisition* atau pengambilan data dari Twitter. Selanjutnya melakukan proses *text mining.* Tahapan di dalam *text mining* menurut [10] adalah:

1. *Data Prepocessing*

Tahapan ini dimana aplikasi melakukan seleksi data teks yang akan diproses. Proses yang dilakukan adalah: *case folding, tokenizing, stemming* dan *tagging. Case folding* adalah proses merubah semua huruf pada sebuah kalimat menjadi huruf kecil dan menghilangkan karakter selain huruf seperti angka, tanda baca dan *Uniform Resources Locator* (URL). *Tokenizing* adalah memotong sebuah kalimat berdasarkan tiap kata yang menyusunnya. Sedangkan *Stemming* adalah merubah berbagai kata berimbuhan menjadi kata dasarnya. Dan *Tagging*, yaitu merubah berbagai kata dalam bentuk lampau menjadi kata awalnya untuk teks dengan Bahasa Inggris.

2. *Feature Selection*

Proses yang dilakukan di tahap ini adalah *stopword removal* yaitu menghilangkan kata-kata yang dianggap tidak penting dari sebuah kalimat.

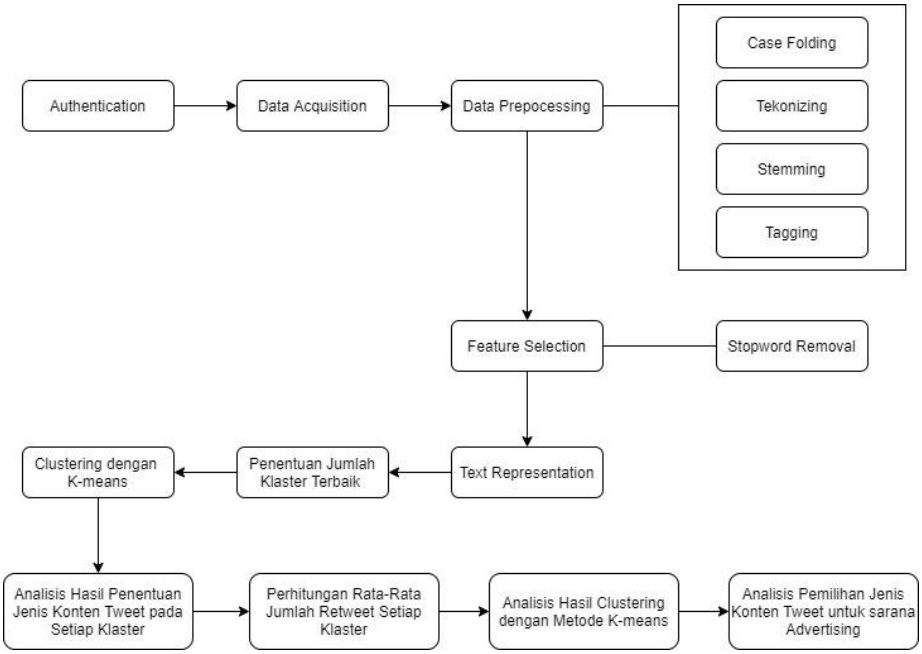
3. *Text Representation*

Tahapan ini merepresentasikan sebuah kalimat sebagai objek dan kata-kata yang menyusunnya sebagai fitur. Dalam tahap ini menggunakan pendekatan *bag of words* atau model ruang vektor dimana sebuah model mempelajari sebuah kosakata dari seluruh dokumen, lalu memodelkan tiap dokumen dengan menghitung jumlah kemunculan setiap kata.

4. *Clustering*

Disini peneliti menggunakan teknik *clustering* dengan menggunakan metode *k-means*. *K-means* adalah salah satu teknik *unsupervised clustering*. *Clustering* dengan metode *k-means* mengelompokkan *n* titik data di dalam *k cluster* dengan meminimasi jarak antara titik data dari pusat *k cluster* secara berulang [12]. Tahapan yang dilakukan adalah [13] : Pertama, memilih jumlah *k cluster* secara acak. Kedua, menghitung jarak antara titik data dengan *centroid* (pusat *cluster*), jarak bisa dihitung menggunakan jarak Euclidean [14] atau pengukuran kosi [15]. Ketiga, titik data ditempatkan pada *cluster* yang memiliki *centroid* dengan jarak minimum. Keempat, setelah semua titik data ditempatkan pada *cluster*, *centroid* dihitung ulang. Terakhir, ulangi tahapan kedua sampai keempat hingga tidak menghasilkan *k centroid* yang baru.

Flowchart metode penelitian yang digunakan oleh peneliti ditampilkan pada gambar 2.



Gambar 2. *Flowchart* Metode Penelitian

4 HASIL DAN PEMBAHASAN

Algoritma proses *text mining* dirancang untuk melakukan *clustering* data *tweets* Gojek Indonesia. Berikut adalah tahapan proses *text mining* yang telah dirancang:

A. *Authentication*

Pada saat bergabung dengan Twitter API, akan mendapatkan sebuah kode berupa *consumer key*, *consumer secret*, dan *access token*, *access token secret* yang hanya bisa di *generated* satu kali. Kode yang peneliti dapat dari Twitter API ditampilkan pada tabel 1.

Tabel 1. Kode Yang Didapat Dari Twitter API

Consumer Key	vNS8jrAaZFQ2bxl03y3ODJclo
Consumer Secret	WlZoXlo0s4LwZ60uhA3QEmKG1bnislIOS9uuZHUg6j70QpPfgs
Acess Token	1262217708998815744-U0agi1BQ7yKh0FkJttcZBJsoeovCgo
Acces Token Secret	C4h9Ho6kgDbZJFpjaGrLfWnL22YFLK2zvASsyRviPIdWD

Kode di atas digunakan untuk proses integrasi antara Twitter API dengan R, dimana proses integrasi dilakukan dengan menggunakan fungsi: 'setup_Twitter_auth (consumer_key, consumer_secret, access_token, access_token_secret)'

B. *Data Acquisition*

Pada tabel 2, adalah contoh data *tweets* yang ditampilkan dari *timeline* Twitter Gojek Indonesia dimana pengambilan data dilakukan secara *real time* menggunakan *package* TwitteR dengan *syntax* yaitu *Tweet* <-userTimeline ("GojekIndonesia", n = 3200, excludeReplies= TRUE). Data yang di *crawling* menghasilkan 84 *tweets* per tanggal 5 juni 2020.

Tabel 2. Contoh Data *Tweets* Yang Ditampilkan Dari *Timeline* Twitter Gojek Indonesia

Nomer *Tweet*	Teks *Tweet*	Tanggal *Tweet*	Jumlah *Retweets/ Comments/Favorites*
1	Cus gaes... pesen makanan yang kamu mau di GoFood sekarang, biar dapet DISKON SAMPE 70% dari promo #HARKULNASGOFOOD! Nikmatin promonya #dirumahaja mulai dari 1 April sampai 5 Mei 2020 ;) *https://t.co/6oQSfi8RFP*	2020- 04-02 04:57:34	*Retweets* = 10 *Comments* = 13 *Favorites* = 41
2	Gaes, ada kabar baik nih. Sekarang, kamu bisa belanja sembako murah buat kebutuhan puasa lewat aplikasi Gojek loh . Iya, ini aku kerja sama dengan @kementan gitu. Cek cara belanja di "Pasar Mitra Tani" via GoFood di video ini ya. Foto: dok. Badan Ketahanan Pangan *https://t.co/aFn0NqpJiT*	2020-04-23 11:44:34	*Retweets* = 38 *Comments* = 16 *Favorites* = 52

C. *Case Folding*

Pada proses *case folding*, dilakukan beberapa fungsi seperti yang ditampilkan pada tabel 3. Untuk contoh hasil proses *case folding* ditampilkan pada tabel 4.

Tabel 3. Fungsi Di Dalam *Case Folding*

No	Fungsi	*Syntax*
1	Ubah huruf kapital menjadi huruf kecil	corpusdokclean <- tm_map(corpusdok, content_transformer(tolower))
2	Hapus Angka	corpusdokclean <- tm_map(corpusdokclean, content_transformer(removeNumbers))
3	Hapus URL	removeURL <- function(x) gsub("http[^[:space:]]*", "",x) corpusdokclean <- tm_map(corpusdokclean, content_transformer(removeURL))
4	Hapus Mention	remove.mention <- function(x) gsub("@\\S+", "", x) corpusdokclean <- tm_map(corpusdokclean, remove.mention)
5	Hapus Hashtag	remove.hashtag <- function(x) gsub("#\\S+", "", x) corpusdokclean <- tm_map(corpusdokclean, remove.hashtag)
6	Hapus Tanda Baca	corpusdokclean<- tm_map(corpusdokclean,content_transformer(removePunctuation))

Tabel 4. Contoh Hasil Proses *Case Folding*

Nomor *Tweet*	Teks *Tweet* Hasil *Case Folding*
1	cus gaes pesen makanan yang kamu mau di gofood sekarang biar dapet diskon sampe dari promo harkulnas gofood nikmatin promonya dirumahaja mulai dari april sampai mei
2	gaes ada kabar baik nih sekarang kamu bisa belanja sembako murah buat kebutuhan puasa lewat aplikasi gojek loh iya ini aku kerja sama dengan kementan gitu cek cara belanja di pasar mitra tani via gofood di video ini ya foto dok badan ketahanan pangan

D. *Tokenizing*

Pada proses *tokenizing*, pemotongan kalimat pada *tweet* berdasarkan tiap kata yang menyusunnya dilakukan dengan menggunakan fungsi: tdm <- TermDocumentMatrix(corpusdokclean, control = list(wordLengths = c(1, Inf))). Contoh hasil proses *tokenizing* ditampilkan pada tabel 5.

Tabel 5. Contoh Hasil Proses *Tokenizing*

Nomor *Tweet*	Hasil *Tokenizing*	Nomer *Tweet*	Hasil *Tokenizing*
1	cus	2	Gaes
	gaes		ada
	pesen		kabar
	makanan		baik
	yang		nih
	kamu		sekarang
	mau		kamu
	di		bisa
	gofood		belanja
	sekarang		sembako
	biar		murah
	dapet		buat
	diskon		kebutuhan
	sampe		puasa
	dari		lewat
	promo		aplikasi
	harkulnas		gojek
	gofood		loh
	nikmatin		iya
	promonya		ini
	dirumahaja		aku
	mulai		kerja
	dari		sama
	april		dengan
	sampai		kementan
	mei		gitu
			cek
			cara
			belanja
			di
			pasar

	mitra tani via gofood di video ini ya foto dok badan ketahanan pangan

E. *Stopword Removal*

Pada proses *stopword removal*, penghilangan kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dari sebuah *tweet* dilakukan dengan menggunakan fungsi: cleanset <- tm_map(corpusdok, removeWords, cStopwordID). Untuk contoh hasil proses *stopword removal* ditampilkan pada tabel 6.

Tabel 6. Contoh Hasil Proses *Stopword Removal*

Nomor *Tweet*	**Hasil *Stopword Removal***	**Nomer *Tweet***	**Hasil *Stopword Removal***
1	cus gaes pesen makanan gofood biar dapet diskon sampe promo harkulnas gofood nikmatin promonya dirumahaja april mei	2	Gaes kabar nih belanja sembako murah kebutuhan puasa aplikasi gojek loh iya kerja kementan gitu cek belanja pasar mitra tani via gofood video ya foto dok badan ketahanan pangan

F. *Text Representation*

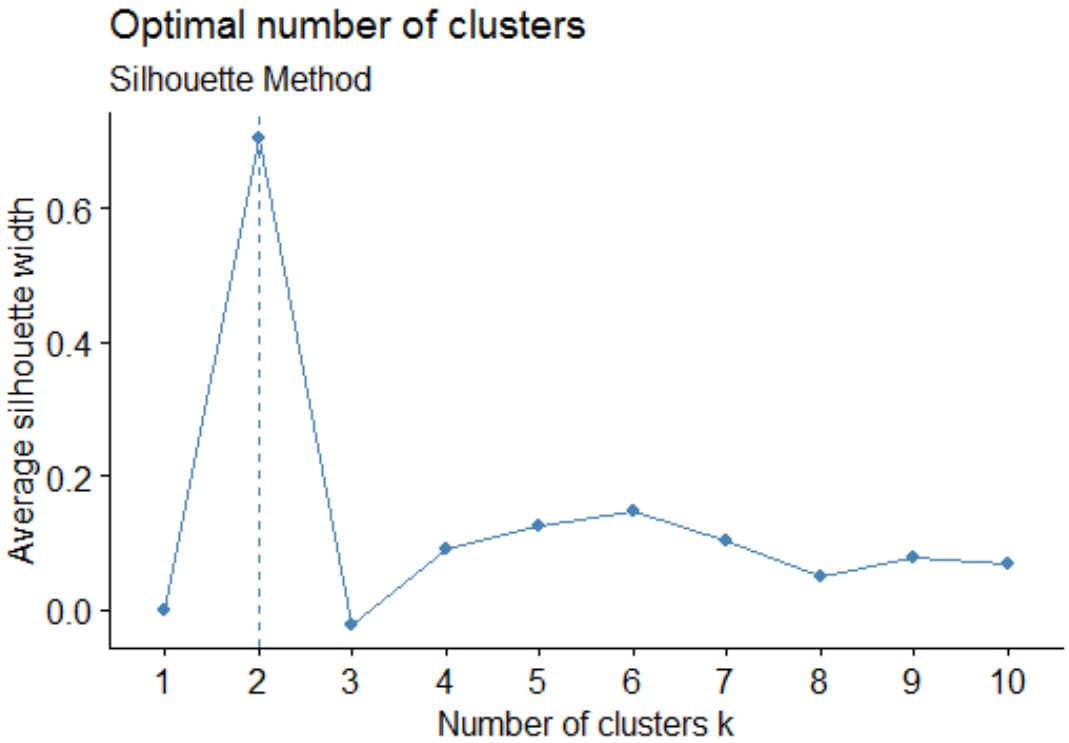
Pada proses *text representation*, perubahan data *tweet* menjadi sebuah matriks dimana baris berupa nomor dari *tweet* dan kolom berupa kata penyusun dari data *tweet* dengan menggunakan fungsi: m <- as.matrix(tdm). Berdasarkan hasil *text representation*, banyaknya seluruh kata yang menyusun 84 *tweet*s dari Gojek Indonesia adalah sebanyak 391 kata. Untuk hasil proses *text representation* ditampilkan pada tabel 7.

Tabel 7. Contoh Hasil Proses *Text Representation*

Nomor *Tweet*	dirumahaja	Gaes	Hobi	ngeliat	ngeshare	penasaran
1	1	1	1	1	1	1
2	0	1	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	0	0	0

G. Penentuan Jumlah Klaster Terbaik

Gambar 3 merupakan nilai *silhouette coefficient* yang dilakukan berdasarkan hasil perhitungan dengan menggunakan fungsi: 'fviz_nbclust (tdm, kmeans, method = "silhouette", k.max=10)'. Perhitungan nilai *silhouette coefficient* dilakukan untuk k = 1 sampai k = 10 dan didapat k terbaik adalah k = 2. Hasil perhitungan k terbaik digunakan untuk proses *clustering* dengan *k-means*.



Gambar 3. Nilai *Silhouette Coefficient* Pada Jumlah Klaster Sebanyak 1 Sampai 10

H. *Clustering* dengan *K-means*

Pada proses *clustering* dengan *k-means* dengan menggunakan fungsi: 'kmeans_i = kmeans (tdm, 2, 100)' dengan jumlah klaster sebanyak 2 klaster dan jumlah iterasi sebanyak 100 iterasi. Untuk hasil *clustering* dengan *k-means* ditampilkan pada tabel 8.

Tabel 8. Hasil *Clustering* Dengan *K-Means*

Nomor Klaster	Nomor *Tweet*	Jumlah *Tweets*
1	9, 25, 38	3
2	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84	81

I. Analisis Hasil Penentuan Jenis Konten *Tweet* pada Setiap Klaster

Proses pemilihan kata yang paling sering muncul pada masing-masing klaster dilakukan dengan menggunakan fungsi: freq <- rowSums(tdm) freq <- subset (freq, freq>=3) freq. Untuk hasil dari proses tersebut ditampilkan pada tabel 9.

Tabel 9. Kata Paling Sering Muncul Dan Jenis Konten *Tweet* Pada Setiap Klaster

Nomor Klaster	Kata Paling Sering Muncul	Jenis Konten *Tweet*
1	gojek, selamat, driver, konser, ramadan, semangat, indonesia	*Rewarding*, konser, ramadan
2	gaes, ebadah, newprofilpic, template, judul, ikutan, komik, cus, kuis, merapat, *tweet*, musafir, dialog	Ajakan, produk gojek, dan kuis

J. Perhitungan Rata-Rata Jumlah *Retweet* Setiap Klaster

Pada tabel 10, merupakan hasil dari proses perhitungan rata-rata jumlah *retweet* tiap klaster dilakukan terhadap hasil dari *clustering* dengan *k-means*.

Tabel 10. Perhitungan Rata-Rata Jumlah *Retweet* Setiap Klaster

Nomor Klaster	Jumlah *Tweets*	Total *Retweets*	Rata-rata *Retweet*
1	3	112	37,333
2	81	3833	47,320

K. Analisis Hasil *Clustering* dengan Metode *K-means*

Salah satu metode yang digunakan untuk menguji kualitas klaster yang dihasilkan dari proses *clustering* adalah *silhouette coefficient.* Pada tabel 11 ditampilkan hasil dari *perhitungan silhouette coefficient* pada setiap klaster *tweet*.

Tabel 11. Perhitungan *Silhouette Coefficient* Pada Setiap Klaster *Tweet*

Nomor Klaster	Jumlah *Tweets*	Nilai *Shilloutte Coeficient*
1	3	0,0
2	81	0,7

Pada gambar 3, dapat dilihat nilai *shiloutte coefficient* pada 2 klaster. Diketahui sebanyak 1 klaster memiliki nilai positif dan 1 klaster memiliki nilai 0. Nilai positif menunjukkan bahwa sebagian besar anggota pada klaster berada pada klaster yang tepat, nilai 0 menunjukkan bahwa sebagian besar anggota pada klaster berada di antara 2 klaster.

Terdapat beberapa faktor yang membuat hasil *clustering* dengan metode *k-means* pada data *tweet* masih belum optimal, diantaranya adalah sebagai berikut:

1. Batas maksimal karakter yang ada pada sebuah *tweet* adalah 280 karakter, sehingga sebagian besar *tweet* mengandung kata-kata yang berupa singkatan. Contoh: seharusnya kata "sosial media" menjadi "sosmed", "by the way" menjadi "btw". Penggunaan kata singkatan berpengaruh terhadap proses *stopword removal*, dimana kata yang seharusnya dihilangkan menjadi tidak terdeteksi sehingga akan tetap ada pada kalimat *tweet*.
2. Tidak dilakukannya proses *stemming* dan *tagging* berakibat pada adanya kata-kata yang serupa maknanya namun beda dalam penulisan hurufnya. Contoh: kata "cus" dengan "cuss". Yang menyebabkan kata ini berada di klaster yang berbeda padahal memiliki arti atau makna yang sama. Selain itu, masih terdapat kata yang berimbuhan. Contoh: "kebutuhan".

L. Analisis Pemilihan Jenis Konten *Tweet* untuk Sarana *Advertising*

Berdasarkan hasil perhitungan jumlah *retweet* pada tiap klaster, didapatkan bahwa jenis konten pada klaster yang memiliki jumlah *retweet* yang tinggi diantaranya tentang ajakan mengikuti kuis dan perkenalan produk dari Gojek Indonesia. Sedangkan jenis konten pada klaster yang memiliki jumlah *retweet* yang rendah mengenai *rewarding*, konser dan Ramadhan.

Hasil yang didapatkan menunjukkan bahwa pelanggan dari Gojek Indonesia lebih tertarik pada program kuis dibandingkan dengan hal yang membahas mengenai *rewarding*, konser dan Ramadhan. Selain itu untuk memperoleh respon dari *followers* Gojek Indonesia, *tweet* yang dibuat oleh Gojek Indonesia sebaiknya berisi hal yang sedang hangat diperbincangkan di masyarakat.

5 KESIMPULAN

Penerapan algoritma proses *text mining* untuk melakukan *clustering* dengan metode *k-means* pada data *tweets* Gojek Indonesia menghasilkan 2 klaster optimal. Berdasarkan perhitungan rata-rata jumlah *retweet* pada tiap klaster, didapatkan bahwa jenis konten pada klaster yang memiliki jumlah *retweet* terbanyak yaitu mengenai program kuis, dan perkenalan produk milik Gojek Indonesia. Pelaku bisnis Gojek Indonesia dapat mengetahui jenis konten *tweet* yang banyak dilakukan *retweet* dan *favorite* oleh *followers*-nya sehingga dapat menggunakan jenis konten *tweet* tersebut sebagai sarana untuk melakukan *advertising* kepada pengguna Twitter. Untuk penelitian selanjutnya, sebaiknya dilakukan proses *stemming* dan *tagging* agar hasil *clustering* yang didapatkan lebih optimal.

REFERENSI

[1] D. Mccorkle And J. Payan, "*Using Twitter In The Marketing And Advertising Classroom To Develop Skills For Social Media Marketing And Personal Branding*," *J. Advert. Educ.*, vol. 21, no. 1, pp. 33–43, May 2017, Doi: 10.1177/109804821702100107.

[2] D. M. Carpenter, J. W. Robertson, M. E. Johnson, And S. Blum, "S*ocial Media Analytics In Education: What Is It, How Is It Useful, And What Does It Tell Us About How Schools Are Discussed In Social Media?," J. Sch. Public Relat.*, vol. 35, no. 1, pp. 7–43, Jan. 2014, Doi: 10.3138/Jspr.35.1.7.

[3] "*Leading Countries Based On Number Of Twitter Users As Of April 2020*," available : https://www.statista.com/statistics/242606/number-of-active-Twitter-users-in-selected-countries/, [Diakses : 18 mei 2020].

[4] D. S. Indraloka And B. Santosa, "Penerapan *Text Mining* Untuk Melakukan *Clustering* Data *Tweet* Shopee Indonesia," *J. Sains Dan Seni Its.*, vol. 6, no. 2, pp. a51–a56, Sep. 2017, Doi: 10.12962/J23373520.V6i2.24419.

[5] R. K. Putri And B. Warsito, "Implementasi Algoritma Modified Gustafson-Kessel Untuk *Clustering Tweet* Pada Akun Twitter Lazada Indonesia," vol. 8, no. 3, p. 11, 2019.

[6] S. Dodi, and S. Iin, " Analisis Sentimen Masyarakat Terhadap data *Tweet* Lazada Menggunakan *Text Mining* Dan Algoritma *Naive Bayes Classifier," Bina Darma Conference on Computer Science.*, e-ISSN: 2685-2683p-ISSN: 2685-2675.

[7] 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, p. 189, May 2018, Doi: 10.33795/Jip.V4i3.205.
[8] I. Putri and Irhamah, "*Text Clustering* pada Akun Twitter Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* dan *K-Means," Jurnal Sains Dan Seni ITS.*, vol. 8, no. 2, 2019.
[9] B. Lantz, "*Machine Learning With R: Learn How To Use R To Apply Powerful Machine Learning Methods And Gain An Insight Into Real-World Applications*," *Birmingham: Packt Publ.*, 2013.
[10] R. Feldman And J. Sanger, "*The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data*," *Cambridge; New York: Cambridge University Press.*, 2007.
[11] Hartanto, " *Text Mining* Dan Sentimen Analisis Twitter Pada Gerakan LGBT", 2017.
[12] J. Kogan, M. Teboulle, And C. Nicholas, "*Data Driven Similarity Measures For K-Means Like Clustering Algorithms*," *Inf. Retr.*, vol. 8, no. 2, pp. 331–349, Apr. 2005, Doi: 10.1007/S10791-005-5666-8.
[13] S. Ahuja And G. Dubey, "*Clustering And Sentiment Analysis On Twitter Data*," p. 5, 2017.
[14] J. Macqueen, "*Some Methods For Classification And Analysis Of Multivariate Observations," Multivar. Obs.*, p. 17.
[15] L. Sahu And B. R. Mohan, "*An Improved K-Means Algorithm Using Modified Cosine Distance Measure For Document Clustering Using Mahout With Hadoop*," *In 2014 9th International Conference On Industrial And Information Systems (Iciis), Gwalior, India.*, pp.1-5, Dec. 2014, Doi: 10.1109/Iciinfs.2014.7036661.