

# Penerapan Algoritma *Fuzzy Gustafson-Kessel* Untuk *Clustering Tweets Mention Akun Go-Jek dan Grab Indonesia*

Firda Ariani Alim Putri<sup>1</sup>, Naim Rochmawati<sup>2</sup>

<sup>1,2</sup>Jurusan Teknik Informatika/Teknik Informatika, Universitas Negeri Surabaya

<sup>1</sup>[firdaputri16051204031@mhs.unesa.ac.id](mailto:firdaputri16051204031@mhs.unesa.ac.id)

<sup>3</sup>[naimrochmawati@unesa.ac.id](mailto:naimrochmawati@unesa.ac.id)

**Abstrak**—Kepopuleran Twitter yang tinggi bisa digunakan untuk berbagai keperluan dalam banyak aspek, seperti untuk kampanye politik, alat pembelajaran, maupun sebagai media komunikasi darurat. Selain itu, tidak sedikit perusahaan yang memanfaatkan Twitter sebagai alat *advertising*, menyebarkan informasi dan promosi. Salah satunya adalah Go-Jek dan Grab Indonesia. Sebagai pelaku bisnis yang memiliki banyak pengikut (*followers*), dituntut untuk memahami jenis konten yang mendapat respon positif dari *followers* agar dapat menentukan strategi pemasaran yang tepat. Selain itu perlu memahami keluhan yang sering dibicarakan para pengikut (*followers*) dalam sebuah *tweet*, sehingga dapat memperbaiki sistem menjadi lebih baik lagi. *Clustering data tweet mention akun @gojekindonesia dan @GrabID menggunakan algoritma Fuzzy Gustafson-Kessel dan indeks validitas Xie-Beni menghasilkan masing-masing 2 cluster optimum. Wordcloud cluster data tweets mention akun @gojekindonesia berisikan keluhan pengguna Go-Jek, sedangkan wordcloud cluster data tweets mention akun @GrabID berisikan ajakan belanja dan promo. Hasil pengujian metode Fuzzy Gustafson-Kessel pada data tweets @gojekindonesia menghasilkan accuracy sebesar 89%, precision sebesar 87%, recall sebesar 92%, dan F-Measure sebesar 89%. Kemudian pada data tweets @GrabID menghasilkan accuracy sebesar 73%, precision sebesar 65%, recall sebesar 76%, dan F-Measure sebesar 70%.*

**Kata Kunci**— *Fuzzy Gustafson Kessel, Index Xie Beni, Clustering, Twitter, Gojek, Grab.*

## I. PENDAHULUAN

Perkembangan internet sebagai *new media (the second media age)* menjadi periode baru dimana teknologi interaktif dan komunikasi jaringan khususnya dunia maya akan mengubah masyarakat [1]. Dengan bertumbuhnya internet yang semakin cepat, mengenalkan kita pada sosial media. Salah satunya adalah Twitter.

Twitter merupakan media sosial yang populer dikalangan masyarakat Indonesia. Menurut Hootsuite: Indonesian Digital

Report 2019, Twitter menduduki peringkat ke-6 sebagai *platform* media sosial yang paling aktif di Indonesia dengan presentase 52%.

Kepopuleran Twitter yang tinggi bisa digunakan untuk berbagai keperluan dalam banyak aspek, seperti untuk kampanye politik, alat pembelajaran, maupun sebagai media komunikasi darurat [2]. Selain itu, tidak sedikit perusahaan yang memanfaatkan Twitter sebagai alat *advertising*, menyebarkan informasi dan promosi. Salah satunya adalah perusahaan yang bergerak di bidang jasa transportasi, yaitu Go-Jek dan Grab Indonesia.

PT. Gojek Indonesia menduduki peringkat pertama sebagai aplikasi penyedia jasa transportasi berdasarkan survei Alvara *Research Center* tentang Perilaku dan Preferensi Konsumen Milenial Indonesia terhadap Aplikasi *E-Commerce* 2019. Go-Jek memiliki 18 layanan di aplikasi, dengan jumlah pengikut di Twitter lebih dari 941rb, pengikut di Instagram sebanyak 933rb, dan 1,3 juta pengikut di Facebook.

Jika Go-Jek lebih terkenal di Indonesia, maka Grab merupakan layanan transportasi terpopuler di Asia Tenggara yang saat ini tersebar di 8 negara dengan 140 juta pengguna dan mitra yang tersebar di Asia Tenggara. Memiliki 13 layanan diaplikasi, dengan jumlah pengikut di Twitter lebih dari 495rb, pengikut di Instagram lebih dari 1 juta, dan 6,6 juta pengikut di Facebook.

Tak sedikit beberapa pengguna Twitter yang juga pengguna jasa transportasi *online* mengeluhkan beberapa kendala yang dialami kemudian memberikan kritik serta saran melalui sebuah *tweet* dan *mention* akun yang bersangkutan agar mendapat respon cepat. Maka sebagai pelaku bisnis yang memiliki banyak pengikut (*followers*), dituntut untuk memahami jenis konten yang mendapat respon positif dari *followers* agar dapat menentukan strategi pemasaran yang tepat. Selain itu perlu memahami keluhan yang sering dibicarakan para pengikut (*followers*) dalam sebuah *tweet*, sehingga dapat memperbaiki sistem menjadi lebih baik lagi.

*Clustering* merupakan proses pengelompokkan data ke dalam beberapa *cluster* atau kelompok, sehingga data dalam satu kelompok memiliki tingkat kemiripan maksimum dan

antar *cluster* memiliki kemiripan minimum [3]. Metode *clustering* yang proses pengelompokannya fleksibel mengikuti bentuk data adalah Algoritma Gustafson-Kessel. *Fuzzy Gustafson-Kessel* merupakan salah satu contoh *clustering* yang merupakan pengembangan dari *Fuzzy C-Means*. [4]

*Fuzzy Gustafson-Kessel* merupakan cara untuk mengelompokan data yang keberadaan data dalam suatu kelompok ditentukan oleh nilai keanggotaan dengan mengubah fungsi perhitungan jarak menjadi fungsi jarak adaptif (*adaptive distance norm*) yang selalu diperbaharui pada setiap iterasi dengan menggunakan matriks *fuzzy covariance*. [5] [6]

Algoritma *Gustafson-Kessel* menggunakan fungsi jarak mahalalanobis sehingga lebih bisa menyesuaikan bentuk geometris pada sebuah himpunan data, tidak seperti *Fuzzy C-Means* yang mengasumsikan bentuk geometris suatu *cluster* adalah bulat sempurna. [7]

Dalam analisis *cluster* diperlukan suatu indeks untuk mengetahui banyaknya *cluster optimum* yang cocok pada suatu penelitian. Salah satu teknik validasi *cluster* yaitu menggunakan indeks Xie-Beni (XB). Xie-Beni memanfaatkan jarak antar *cluster* dan jumlah data untuk menentukan nilai indeks XB. Dimana semakin kecil nilai indeks XB jumlah *cluster*, maka jumlah *cluster* tersebut semakin optimal. [8]

Rekomendasi penggunaan indeks XB dijelaskan dalam penelitian Duo dkk (2007) yang menyebutkan bahwa indeks Xie-Beni memiliki ketepatan yang tinggi untuk digunakan sebagai kriteria dalam menentukan kelompok optimum pada metode *hard partition* seperti *K-Means cluster* maupun pada FCM. [9]

Tujuan dari penelitian ini adalah untuk mengetahui konten yang menarik respon positif dan topik apa saja yang banyak dibicarakan pengguna jasa transportasi online. Hal ini dapat membantu perusahaan terkait untuk memperbaiki layanan dan kualitas perusahaan menjadi lebih baik.

Dalam penelitian ini penulis melakukan pengelompokan data *tweets mention* akun @gojekindonesia dan @GrabID menggunakan Metode *Fuzzy Gustafson-Kessel*. Pada penelitian ini data *tweets* akan diproses menggunakan bahasa pemrograman Python.

## II. PENELITIAN TERKAIT

Sebagai tolok ukur keberhasilan metode yang digunakan agar tidak terjadi plagiasi, perlu dilakukan analisis penelitian terdahulu. Seperti halnya penelitian mengenai *clustering* data Twitter, diantaranya penelitian yang dilakukan Dwi Smaradahana Indraloka dan Budi Santosa pada tahun 2017 tentang penerapan *text mining* untuk menerapkan *clustering* data *tweet* Shopee Indonesia. Metode *Clustering* yang

digunakan adalah *K-means* dan untuk menentukan jumlah *cluster* terbaik menggunakan metode *Silhouette coefficient*. Penelitian ini dilakukan untuk mengetahui jenis konten apa yang disukai pengguna Twitter pada akun Shopee Indonesia. Hasilnya adalah kuis berhadiah, ulang tahun Shopee, dan hobi, kuis, dan *lifestyle* yang memiliki jumlah retweet tinggi. [10]

Pada tahun 2017, Nabila Safina dan Aris Marjuni melakukan penelitian analisis sentimen pada Twitter terhadap jasa transportasi *online* di Indonesia menggunakan metode *Support Vector Machine*. Penelitian ini bertujuan untuk melihat respon masyarakat terhadap jasa transportasi Gojek dan Grab Indonesia. Sebanyak 1000 data twitter yang *mention* akun @gojekindonesia dan @GrabID diklasifikasi menjadi 2 kelas tweet, yaitu positif dan negatif. Proses klasifikasi menunjukkan hasil akurasi sebesar 87%. [11]

Ahmad Mauliyadi M, Hizir Sofyan, dan Muhammad Subianto melakukan penelitian dengan membandingkan metode *Fuzzy C-Means (FCM)* dan *Fuzzy Gustafson-Kessel (FGK)* menggunakan data citra satelit quickbird pada tahun 2013. Penelitian ini menggolongkan data yang tak terdefinisi atau kabur ke dalam kelompok (*cluster*) yang ditentukan berdasarkan derajat keanggotaan yang bernilai antara 0 hingga 1. Kemudian membandingkan metode yang terbaik dengan melihat perbandingan visualisasi dan pengambilan data secara acak kemudian divalidasi dengan data sebenarnya. Hasilnya adalah dengan parameter yang sama, *Fuzzy Gustafson-Kessel* lebih akurat daripada *Fuzzy C-Means*. Hal ini berdasarkan hasil akurasi *Fuzzy Gustafson-Kessel* pengelompokan sebesar 88% sedangkan *Fuzzy C-Means* sebesar 62% dari data sebenarnya. [12]

Di tahun 2015, Laily dkk melakukan penelitian analisis kelompok menggunakan algoritma *Fuzzy C-Means* dan Gustafson-Kessel *Clustering* pada index LQ45. Penelitian ini mengelompokan rasio saham menggunakan algoritma *Fuzzy C-Means* dan Gustafson-Kessel dengan variabel *Earning Per Share (EPS)* dan *Price Earning Ratio (PER)*. Hasil pengelompokan menunjukkan algoritma Gustafson-Kessel memenuhi kriteria fungsi objektif minimum, nilai rasio kecil, dan index Xie dan Beni minimum dibandingkan dengan algoritma *Fuzzy C-Means*. [13]

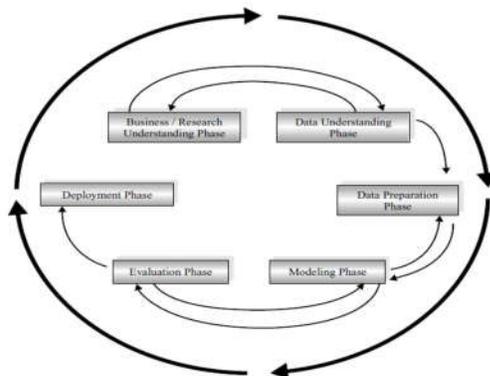
Penelitian selanjutnya oleh Krishna Priya C.B dan Dr. S. Venkateswari tahun 2017 dengan menggunakan algoritma *clustering* yaitu *C-Means* dan Gustafson-Kessel untuk mempelajari penggambaran dari zona pengelolaan pertanian. Variabel yang digunakan sebagai data input antara lain konduktivitas listrik, NPK, luas tanah, dll. Hasilnya adalah Gustafson-Kessel lebih efisien dalam waktu pelaksanaan dimana prosesnya diukur dalam milidetik dibandingkan *C-*

Means, namun Gustafson-Kessel menggunakan memori lebih tinggi dibandingkan dengan prototype dasar. [14]

Kemudian pada tahun 2019 Ratna Kencana Putri, Budi Warsito, dan Mustafid melakukan penelitian menggunakan algoritma *Modified Gustafson-Kessel* untuk *clustering tweets* pada akun Twitter Lazada Indonesia untuk mengetahui konten yang disukai pengguna Lazada berdasarkan jumlah *retweet* dan *like*. Untuk menentukan konten yang disukai pengikut (*followers*), dilakukan perhitungan rata-rata jumlah *retweet* dan jumlah *like* masing-masing *cluster*. Metode *clustering* menggunakan algoritma *Fuzzy Gustafson-Kessel* dengan indeks validitas *Partition Coefficient* (PC) dan *Classification Entropy* (CE). Penelitian ini menghasilkan 3 *cluster tweets* kemudian dilakukan *profiling* masing-masing *cluster* dengan bantuan *wordcloud* sehingga menghasilkan konten yang disukai pengguna Lazada Indonesia. [4]

### III. METODOLOGI PENELITIAN

Pelaksanaan penelitian ini dilakukan dengan menggunakan model CRISP-DM (*Cross Industry Standart Process Model for Data Mining*). Gbr 1 merupakan gambaran alur yang terjadi pada CRISP-DM. (Laroze, 2014)



Gbr 1. Alur CRISP-DM

Penerapan model CRISP-DM dalam penelitian ini adalah sebagai berikut:

#### 1) Business Understanding

Pada tahapan ini diperlukan pemahaman terhadap pentingnya pemanfaatan data tweet, agar bisa digunakan untuk mengetahui konten yang menarik respon positif dan topik apa saja yang banyak dibicarakan pengguna jasa transportasi online. Hal ini bertujuan untuk membantu perusahaan terkait untuk memperbaiki layanan dan kualitas perusahaan menjadi lebih baik.

#### 2) Data Understanding

Data dikumpulkan dengan cara *crawling data tweets* dengan bantuan Twitter API dan Bahasa pemrograman Python untuk proses *clustering*. Data *tweets* diambil dari *tweet* yang melakukan *mention* pada akun @gojekindonesia dan @GrabID selama beberapa bulan terakhir.

#### 3) Data Preparation

Langkah yang dilakukan pada atahap ini adalah pengolahan data. Proses pengolahan data ini terdiri dari *pre-processing*, pembobotan TF-IDF (*Term Frequency Invers Document Frequency*), *clustering*, dan *wordcloud*. Data Twitter berupa teks diambil (*crawling*) kemudian diolah menjadi suatu matriks angka melalui tahap *pre-processing* dan pembobotan TF-IDF melalui persamaan (1):

$$W_{j,i} = \frac{n_{j,i}}{\sum_{j=1}^p n_{j,i}} \cdot \log_2 \frac{D}{d_j} \quad (1)$$

Dengan:

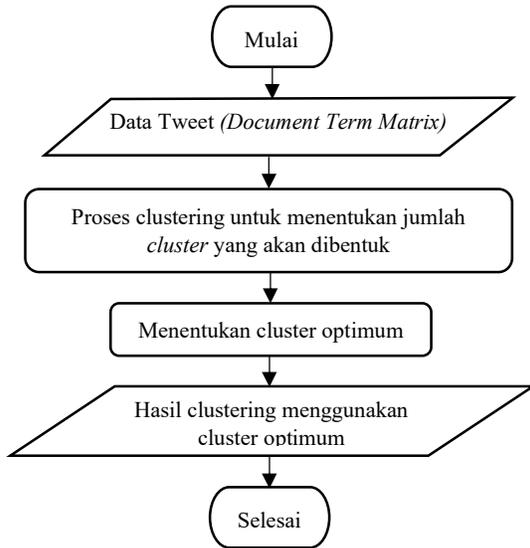
- $W_{j,i}$  = pembobotan TF-IDF untuk *term* ke-j pada dokumen ke-i
- $n_{j,i}$  = jumlah kemunculan *term* ke-j pada dokumen ke-i
- $p$  = banyaknya *term* yang terbentuk
- $\sum_{j=1}^p n_{j,i}$  = jumlah kemunculan seluruh *term* dokumen ke-i
- $D$  = banyaknya dokumen yang dibangkitkan
- $d_j$  = banyaknya dokumen yang mengandung *term* ke-j

Pembobotan TF-IDF atau *Term Frequency Invers Document Frequency* merupakan metode yang digunakan untuk menentukan keterhubungan kata (*term*) pada dokumen dengan memberikan bobot disetiap kata. [15]

Pada proses pembobotan TF-IDF akan menghasilkan *matrix* yang nantinya digunakan dalam proses *clustering* algoritma *Gustafson-Kessel*. Dengan baris pada *matrix* mewakili dokumen (*tweet*) yang digunakan, sedangkan kolom *matrix* berisi kata-kata yang digunakan untuk mengetahui konteks dari *tweet* tersebut (*terms*). [4]

#### 4) Data Modelling

Pada tahap ini dilakukan proses *clustering* untuk menentukan jumlah *cluster* yang akan dibentuk menggunakan data yang sudah diolah menjadi *Document Term Matrix* dengan pembobotan TF-IDF. Kemudian menentukan jumlah *cluster optimum* yang ditentukan menggunakan indeks validitas Xie dan Beni. Dengan alur yang dijelaskan pada Gbr 2.



Gbr 2. Alur Proses Clustering

a) Proses Clustering Algoritma Gustafson-Kessel adalah sebagai berikut: [7]

- Memasukkan data yang ingin dikelompokkan sebagai  $x$  (matriks  $n \times p$ )
- Menentukan jumlah *cluster* yang ingin dibentuk ( $c \geq 2$ ), *weighting exponent* ( $m > 1$ ), iterasi maksimum ( $t_{max}$ ), error terkecil yang diinginkan ( $\epsilon$ ), nilai *threshold* ( $\beta$ ), dan parameter pembobot  $\gamma \in [0,1]$ .
- Membangkitkan bilangan acak  $U_{ik}$ ,  $1 \leq i \leq n$ ;  $1 \leq k \leq c$  sebagai matriks partisi awal  $U_0$  dan menghitung matriks kovarian  $F_0$  dari semua data.

$$\sum_{k=1}^c U_{ik}^m = 1, 1 \leq i \leq n; 1 \leq k \leq c \quad (2)$$

- Menghitung pusat *cluster* ke- $k$ :  $v_k$

$$v_k^{(t)} = \frac{\sum_{i=1}^n (u_{ik}^{(t-1)})^m x_i}{\sum_{i=1}^n (u_{ik}^{(t-1)})^m}, 1 \leq k \leq c \quad (3)$$

Dengan:

- $u_{ik}$  = derajat keanggotaan data ke- $i$  pada *cluster* ke- $k$
- $m$  = pangkat pembobot untuk fungsi keanggotaan *fuzzy*
- $t$  = banyak iterasi
- $n$  = banyak data
- $c$  = banyak *cluster*

- Menghitung matriks kovarian *cluster*

$$F_k = \frac{\sum_{i=1}^n (u_{ik}^{(t-1)})^m (x_i - v_k^{(t)})(x_i - v_k^{(t)})^T}{\sum_{i=1}^n (u_{ik}^{(t-1)})^m}, \quad 1 \leq k \leq c \quad (4)$$

Dengan:

- $x_i$  = vector data ke- $i$
- $v_k$  = pusat *cluster* ke- $k$
- $u_{ik}$  = derajat keanggotaan data ke- $i$  pada *cluster* ke- $k$
- $m$  = pangkat pembobot fungsi keanggotaan *fuzzy*
- $t$  = banyak iterasi
- $n$  = banyak data
- $c$  = banyak *cluster*

Ekstraksi nilai *eigenvectors*  $\phi$  dan *eigenvalues*  $\Lambda$  dari  $F_k^{new}$  yang telah dihitung dengan persamaan berikut:

$$F_k^{new} = (1 - \gamma)F_k + \gamma \det(F_0)^{1/p} \mathbf{I} \quad (5)$$

Dengan:

- $\gamma$  = parameter pengatur bentuk matriks *fuzzy covariance*,  $\gamma \in [0,1]$
- $F_0$  = matriks kovarian seluruh data
- $F_k$  = matriks *fuzzy covariance cluster* ke- $k$  (persamaan 4)
- $p$  = banyak variabel
- $\mathbf{I}$  = matriks identitas

Jika rasio antara nilai *eigenvectors* dan *eigenvalues* melewati ketentuan nilai *threshold*, maka rekonstruksi  $F_k$  dijelaskan sebagai berikut:

$$F_k = \phi \Lambda \phi^{-1} \quad (6)$$

Dengan:

- $\phi$  = vector eigen matriks *fuzzy covariance cluster* ke- $k$

$\Lambda$  = matriks diagonal dari nilai eigen matriks *fuzzy covariance cluster* ke-k

- Menghitung jarak dengan persamaan berikut:  

$$D_{ikA_k}^2 = (\mathbf{x}_i - \mathbf{v}_k^{(t)})^T [\rho_k \det(\mathbf{F}_k)^{1/p} \mathbf{F}_k^{-1}] (\mathbf{x}_i - \mathbf{v}_k^{(t)})$$
(7)

Dengan:

$i = 1, 2, \dots, n$  dan  $k = 1, 2, \dots, c$

$D_{ikA_k}^2$  = jarak data ke-i terhadap pusat *cluster* ke-k dengan *norm inducing matrix*  $A_k$

$\mathbf{x}_i$  = vector data ke-i

$\mathbf{v}_k$  = pusat *cluster* ke-k

$\mathbf{F}_k$  = matriks *fuzzy covariance cluster* ke-k

$\rho_k$  = volume *cluster* ke-k

$\rho$  = banyak variabel

$n$  = banyak data

$c$  = banyak *cluster*

- Memperbarui matriks fungsi keanggotaan  
 Untuk  $1 \leq i \leq n$  jika  $D_{ikA_k}^2 > 0$  untuk  $1 \leq k \leq c$

$$u_{ik}^{(t)} = \left[ \sum_{l=1}^c \left( \frac{D_{ikA_k}^2}{D_{ilA_k}^2} \right)^{\frac{2}{m-1}} \right]^{-1}$$
(8)

Jika tidak, maka

$$u_{ik}^{(t)} = 0 \text{ jika } D_{ikA_k}^2 > 0 \text{ dan } u_{ik}^{(t)} \in [0,1]$$
(9)

dengan  $\sum_k^c = 1$  dan  $u_{ik}^{(t)} = 1$ , sampai

$\|\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}\| < \varepsilon$  atau jika  $t >$  iterasi maksimum.

Nilai *threshold* ( $\beta$ ) biasanya akan ditentukan dalam angka besar, seperti  $10^{15}$ . Untuk menentukan nilai parameter pembobot  $\gamma$  tergantung pada data yang digunakan, beberapa percobaan mungkin perlu dilakukan untuk menentukan nilai  $\gamma$  yang tepat.

#### b) Penentuan Jumlah *Cluster Optimum* (Xie-Beni)

Penentuan jumlah *cluster optimum* menggunakan perhitungan indeks validitas Xie-Beni (Purnamasari, 2014). Indeks Xie-Beni bertujuan untuk mengukur rasio dari total variasi di dalam *cluster* ( $S$ ) dan pemisah antar *cluster* ( $d_{min}$ ). Banyaknya *cluster optimum* diperoleh dengan meminimalkan nilai indeks (Jansen, 2007). Kriteria banyaknya *cluster optimum* diberikan oleh nilai  $XB$  minimum pada lembah pertama (Kuo, 2004). [16]

Rumus Index Xie Beni untuk penentuan *cluster optimum* sebagai berikut: [16]

$$XB(c) = \frac{S}{d_{min}}$$
(10)

dimana,

$$S = \frac{\sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m d_{ik}^2(x_i, v_k)}{n}$$
(11)

$$d_{min} = \min_{i \neq k} d_{ik}^2(v_i, v_k)$$
(12)

Dengan:

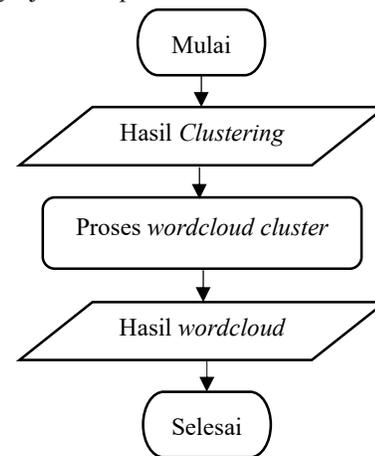
$d_{ik}(x_i, v_k)$  = jarak data terhadap pusat *cluster*

$d_{ik}(v_i, v_k)$  = jarak pusat *cluster*  $v_k$  ke pusat *cluster*  $v_i$

$m$  = pangkat pembobot (*fuzzifier*)

#### c) Proses Interpretasi *cluster*

Untuk menginterpretasi setiap *cluster*, dibantu dengan *wordcloud*. *Wordcloud* digunakan untuk melihat kata-kata yang menyusun masing-masing *cluster*. Dengan alur yang dijelaskan pada Gbr 3.



Gb 3. Alur Proses Interpretasi *Cluster*

## IV. HASIL DAN PEMBAHASAN

Pada penelitian ini, proses awal yang dilakukan adalah *crawling* data *tweet mention* akun @gojekindonesia dan @GrabID. yang diunggah sebelum tanggal 12 Desember 2020 menggunakan *Python*. Data teks yang digunakan dalam penelitian ini adalah *tweets* dengan bahasa Indonesia dari *mention* akun @gojekindonesia dan @GrabID.

### 1) *Crawling* Data

Pada akun @gojekindonesia jumlah *tweets* maksimal yang dapat diambil adalah sebanyak 886 *tweets*. Beberapa hasil *crawling* data *tweets* tertera pada Tabel 1.

TABEL 1  
HASIL CRAWLING DATA TWEETS @GOJEKINDONESIA

No	Hasil Pembobotan TF-IDF					
	cashback	driver	grabfood	...	ovoid	promo
119	0.239	0	0	...	0.720837	0
120	0	0	0	...	0	0

No	Text Tweet	Tanggal Tweet
1.	@gojekindonesia min, mau tanya kok pas mau bayar pake gopay dibilangnya begini terus kenapa yaa? Padahal udah versi terbaru. <a href="https://t.co/IJTA9EgSJ">https://t.co/IJTA9EgSJ</a>	11/12/2020 23:50 PM
2.	@Rtior11 Hai, perihal kendala Kamu silakan memberikan informasi lebih lanjut melalui DM ya. Terima kasih^ali	11/12/2020 23:10 PM
3.	@gojekindonesia tolong admin cek dm saya. Masa kalah sama sebelah fast respondnya\ud83d\ude43	11/12/2020 16:44 PM

Sedangkan pada akun @GrabID jumlah tweets maksimal yang dapat diambil adalah sebanyak 629 tweets. Beberapa hasil crawling data tweets tertera pada Tabel 2.

TABEL 2  
HASIL CRAWLING DATA TWEETS @GRABID

No	Text Tweet	Tanggal Tweet
1.	RT @BerburuSale: Yang ditunggu udah dateng nih, promo BOOM dari @grabid @ovo_id @tokopedia ada lagi!\n\nBelanja #pakeOVOaja nikmati cashback...	11/12/2020 12:56 PM
2.	@Syawar123 Mohon maaf Kak. Saat ini masih dalam pengecekan oleh tim terkait Kak. Mohon menunggu dan melakukan pengecekan secara berkala ya. Terima kasih 🙏 - Budi	11/12/2020 12:55 PM
3.	RT @gigiigiaa: @radenrauf @GrabID @ovo_id @tokopedia Wishlist aku \nBeli skincare\nBeli makeup\nBeli baju\nBeli sepatu\n#BOOM1212	11/12/2020 04:45 AM

## 2) Pengolahan Data

Data tweets diolah menjadi suatu matrix angka melalui tahapan *text pre-processing* dan pembobotan TF-IDF. Pembobotan TF-IDF digunakan untuk melihat kecenderungan kata yang sering muncul pada data tweets. Semua kata tersebut akan menjadi variabel tiap tweet, dengan komponen dari matrix berupa jumlah suatu kata pada setiap tweet. Kemudian matrix angka nantinya digunakan saat proses *clustering*.

Berdasarkan hasil Pembobotan TF-IDF jumlah kata yang menyusun 886 tweets adalah 3.335 kata. Sehingga matrix yang terbentuk adalah 886 tweets sebagai baris dan 3.335 kata sebagai kolom pada matrix. Pembobotan TF-IDF menggunakan persamaan (1) sehingga hasil pembobotan TF-IDF tertera pada Tabel 3.

TABEL 3  
HASIL PEMBOBOTAN TF-IDF @GOJEKINDONESIA

No	Hasil Pembobotan TF-IDF					
	cashback	driver	gofood	...	gopay	promo
250	0.161	0	0	...	0.107	0.156
251	0	0	0	...	0	0

Kemudian pada data tweets @GrabID, jumlah kata yang menyusun 629 tweets adalah 1.924 kata. Sehingga matrix yang terbentuk adalah 629 tweets sebagai baris dan 1.924 kata sebagai kolom. Pembobotan TF-IDF menggunakan persamaan (1) sehingga hasil pembobotan TF-IDF tertera pada Tabel 4.

TABEL 4  
HASIL PEMBOBOTAN TF-IDF @GRABID

## 3) Clustering Data

Data yang digunakan adalah hasil dari pembobotan TF-IDF. Proses *clustering* data @gojekindonesia dan @GrabID dilakukan sebanyak 5 kali dengan jumlah cluster sebanyak 2 cluster sampai dengan 6 cluster.

Hasil yang didapat pada data @gojekindonesia yaitu jumlah cluster yang memiliki anggota kelompok sebanyak 6 cluster. Sehingga dapat disimpulkan bahwa tweets mention akun @gojekindonesia dapat dikelompokkan dengan jumlah 6 cluster.

Maka dilakukan pengukuran Index Xie Beni untuk menentukan cluster optimum data @gojekindonesia dengan jumlah 2 hingga 6 cluster. Menggunakan persamaan (10), yaitu jarak data dari pusat cluster dibagi dengan jarak pusat cluster. Tabel 5 merupakan hasil XB untuk setiap jumlah cluster.

TABEL 5  
HASIL INDEX XIE-BENI DATA @GOJEKINDONESIA

Jumlah Cluster	Nilai Xie-Beni
2	24878994.7433
3	303340719.3667
4	499246372.1357
5	157980240.5833
6	116222476.9064

Maka dapat disimpulkan bahwa jumlah cluster optimum diberikan untuk jumlah cluster sebanyak 2 cluster karena memiliki nilai XB minimum. Hasil clustering dengan jumlah 2 cluster dapat dilihat pada Tabel 6.

TABEL 6  
HASIL CLUSTERING DATA TWEETS @GOJEKINDONESIA MENGGUNAKAN ALGORITMA GUSTAFSON-KESSEL

Cluster ke-	Nomor Anggota Tweet	Jumlah Anggota
1	1, 2, 4, 8, 9, 10, 12, 14, 16, 17, 19, 26, 27, 28, 29, 30, 31, 32, 34, 37, 39, 42, 43, ..., 875, 879, 880, 881, 882, 884	560



bahwa *cluster 2* adalah *tweet* yang berisikan promo belanja *online* di salah satu situs belanja *online* bersama Grab.



Gb. 7 Wordcloud Cluster 2 Grab

### 5) Hasil Pengujian

Pada proses ini dilakukan pengujian metode *Fuzzy Gustafson-Kessel* terhadap 100 *tweets* secara acak dari masing-masing data dan menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F-Measure* seperti pada tabel 9.

TABEL 9  
HASIL PENGUJIAN METODE FUZZY GUSTAFSON-KESSEL

Data	Accuracy	Precision	Recall	F-Measure
@gojekindonesia	89%	87%	92%	89%
@GrabID	73%	65%	76%	70%

## V. KESIMPULAN

Berdasarkan hasil pengujian, kesimpulan yang didapat dari penerapan algoritma *Gustafson-Kessel* untuk *clustering tweets* akun @gojekindonesia dan @GrabID adalah:

1. Penerapan Algoritma *Gustafson-Kessel* untuk *clustering tweets mention* akun @gojekindonesia menghasilkan 2 *cluster tweets* mengenai keluhan terhadap sistem di aplikasi. Kemudian *clustering tweets mention* akun @GrabID menghasilkan 2 *cluster tweets* mengenai ajakan belanja menggunakan pembayaran ovo dan promo.
2. Hasil pengujian metode *Fuzzy Gustafson-Kessel* pada data *tweets* @gojekindonesia menghasilkan *accuracy* sebesar 89%, *precision* sebesar 87%, *recall* sebesar 92%, dan *F-Measure* sebesar 89%. Kemudian pada data *tweets* @GrabID menghasilkan *accuracy* sebesar 73%, *precision* sebesar 65%, *recall* sebesar 76%, dan *F-Measure* sebesar 70%.

## VI. SARAN

Data *tweet* yang diperoleh sebagian besar merupakan *tweet mention* terhadap akun @gojekindonesia dan @GrabID, hasil *wordcloud* dari setiap *cluster* cenderung sama sehingga sulit untuk ditarik kesimpulan pada proses *wordcloud*.

Untuk penelitian selanjutnya disarankan untuk menggunakan data *tweets* yang lebih spesifik serta menggunakan metode *clustering* yang lebih efisien dan akurat.

## UCAPAN TERIMA KASIH

Puji syukur kepada Allah SWT yang telah melimpahkan rahmat, kesehatan dan ilmu kepada penulis sehingga bias menyelesaikan penelitian ini. Serta kepada pihak yang telah membantu penulis sampai terselesaikannya artikel ini.

## REFERENSI

- [1] Littlejohn. dkk, "Teori Komunikasi," Jakarta, Salemba Humanika, 2009, p. 413.
- [2] Naim Rochmawati & S.C. Wibawa, "Opinion Analysis on Rohingya using Twitter Data," *IOP Conf. Series: Materials Science and Engineering*, 2018.
- [3] Tan. dkk, *Introduction to Data Mining*, Boston: Pearson Education, 2006.
- [4] Ratna Kencana Putri. dkk, "Implementasi Algoritma Modified Gustafson-Kessel Untuk Clustering Tweets Pada Akun Twitter Lazada Indonesia," *Jurnal Gaussian*, vol. 8 No.3, pp. 285-295, 2019.
- [5] Gustafson D & Kessel W, "Fuzzy Clustering with A Fuzzy Covariance Matrix," *Conference Paper in Proceeding of the IEEE Conference on Decision and Control*, 1978.
- [6] Dwijayani. M, "Pengelompokan Entitas Audit Pemerintahan Daerah di BPK RI Perwakilan Jabar Dengan Metode Clustering," *Journal Program Magister Management Teknologi ITS*, 2010.
- [7] Babuska R. dkk, "Improved Covariance Estimation for Gustafson-Kessel Clustering," *In Proceeding of the IEEE International Conference on Fuzzy Systems*, pp. 1081-1085, 2002.
- [8] Lathief M. F, "Implementasi Metode Xie Beni untuk Validasi Cluster Pada Fuzzy C-Means Studi Kasus: Indikator Kesehatan Masyarakat, UII Yogyakarta," 2016.
- [9] Pravitari. dkk, "Penentuan Banyak Kelompok dalam Fuzzy C-Means Cluster Berdasarkan Proporsi Eigen Value Dari Matriks Similarity dan Indeks XB (Xie dan Beni)," 2011.
- [10] Santosa. dkk, "Penerapan Text Mining untuk Menerapkan Clustering Data Tweet Shopee Indonesia,"

*Jurnal Sains dan Seni ITS*, Vol. 16 No.2, pp. A51-A56, 2017.

- [11] Marjuni. dkk, ". Analisis Sentimen Pada Twitter Terhadap Jasa Transportasi Online di Indonesia Dengan Metode Support Vector Mechine.," *Draft Jurnal FIK UDINUS*, 2017.
- [12] Mauliyadi A. dkk, "Perbandingan Metode Fuzzy C-Means dan Fuzzy Gustafson-Kessel Menggunakan Data Citra Satelit Quickbird," *Jurnal Transenden*, pp. 1-5, 2013.
- [13] Laili Rahmatika. Suparti & Safitri, "Analisis Kelompok Dengan Algoritma Fuzzy C-Means dan Gustafson Kessel Clustering Pada Index LQ45," *Jurnal Gaussian*, vol. 4 No.3, pp. 543-552, 2015.
- [14] Venkateswari. dkk, "Application of Gustafson-Kessel-Like Clustering Algorithm in Delineation of Management Zones in Precision Agriculture," *International Journal of Applied Agricultural Research*, vol. 12 No.3, pp. 279-293, 2017.
- [15] Fitri & Meisya, "Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi TF-IDF Untuk Pencarian Dokumen Berbahasa Indonesia," 2013.
- [16] Desy Rahmawati Ningrat. dkk, "Analisis Cluster Dengan Algoritma K-Means dan Fuzzy C-Means Clustering Untuk Pengelompokkan Data Obligasi Korporasi," *Jurnal Gaussian*, vol. 5 No.4, 2016.