ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 8, Nomor 3, Tahun 2019, Halaman 285 - 295
Online di: http://ejournal3.undip.ac.id/index.php/gaussian

IMPLEMENTASI ALGORITMA *MODIFIED* GUSTAFSON-KESSEL UNTUK *CLUSTERING TWEETS* PADA AKUN TWITTER LAZADA INDONESIA

Ratna Kencana Putri[1], Budi Warsito[2], Mustafid[3]
[1,2,3]Departemen Statistika FSM Universitas Diponegoro
budiwrst2@gmail.com

ABSTRACT

Online social media is a new kind of media which is steadily growing and has become publicly popular. Due to its ability to spread informations rapidly and its easiness to access for internet users, social media provides new alternative to conduct advertising and product segmentation. Twitter is one of the most favored social media with 19.5 million users in Indonesia to the date. In this research, the application of text mining to cluster tweets from the @LazadaID Twitter account is done using the Modified Gustafson-Kessel clustering algorithm. The clustering process is executed five times with the number of cluster starts from two to six cluster. The results of this research indicate that the optimum number of clusters formed based on the Partition Coefficient and Classification Entropy validation index are three clusters. Those three clusters are tweets containing electronic stuff offers, discounts, and prize quizes. Tweets with the most retweets and likes are prize quiz tweets. PT Lazada Indonesia could use this kind of tweet to conduct advertising on social media Twitter because the prize quiz tweets are liked by the @LazadaID Twitter account followers.

Keywords: Twitter, advertising, Lazada Indonesia, Gustafson-Kessel Clustering algorithm, validation index.

1. PENDAHULUAN

Pada era digital ini, berbagai macam teknologi komunikasi hadir di kehidupan manusia untuk memperoleh informasi. Perkembangan internet sebagai media baru (*the second media age*) menandai periode baru dimana teknologi interaktif dan komunikasi jaringan khususnya dunia maya akan mengubah masyarakat [8]. Internet memberikan kemudahan bagi para penggunanya untuk mengakses informasi dengan sistem *online* yang dapat diakses kapanpun dan dimanapun. Media sosial *online* telah menjadi suatu media baru yang semakin berkembang dan populer di lingkungan masyarakat dan secara umum digunakan untuk mengekspresikan diri dan menyampaikan pendapat pengguna terhadap segala hal secara bebas.

Twitter merupakan salah satu media sosial yang paling populer di Indonesia dengan 19,5 juta pengguna di Indonesia dari total 500 juta pengguna global [10]. Setiap pengguna Twitter dapat dengan bebas membuat *tweets* dengan kalimat apapun yang menggambarkan kehidupan, pendapat, ataupun kejadian tertentu. Dengan fasilitas tersebut maka banyak perusahaan yang akhirnya menggunakan Twitter sebagai alat *advertising* serta media untuk menyebarkan informasi. Salah satu perusahaan yang menggunakan Twitter sebagai alat promosi adalah PT Lazada Indonesia dengan nama akun twitter @LazadaID.

PT Lazada Indonesia menduduki peringkat pertama *e-commerce* paling top di Indonesia dengan jumlah kunjungan ke laman lebih dari 117 juta *visit*, 364 ribu pengikut di Twitter, 1 juta pengikut di Instagram, dan 22,7 juta pengikut di Facebook [1]. Lazada menggunakan akun Twitter tersebut untuk memberikan informasi mengenai produk-produk yang dijual, promo yang sedang ditawarkan, serta penawaran-penawaran lain untuk menarik

minat pembeli agar melakukan transaksi di Lazada. Pelaku bisnis harus mampu memahami jenis konten yang mendapat respon positif dari *followers*, agar dapat menentukan strategi pemasaran yang tepat.

Metode *text mining* dapat digunakan untuk menganalisa data pada Twitter karena sebagian besar informasi yang tersedia dalam Twitter disimpan dalam bentuk data teks [6]. Salah satu tahapan lanjutan dari metode *text mining* yaitu *clustering* untuk menentukan pola dan struktur yang menarik dari sebuah data yang berjumlah sangat besar dengan latar belakang pengetahuan yang sedikit [13]. Metode *clustering* yang saat proses pengelompokannya fleksibel mengikuti bentuk data adalah Algoritma Gustafson-Kessel. Dalam analisis *cluster* memerlukan suatu indeks untuk mengetahui banyaknya *cluster* optimum yang cocok pada suatu penelitian [9]. Salah satu, indeks validitas yang cocok untuk *fuzzy clustering* adalah indeks validitas *Partition Coefficient* (PC) dan *Classification Entropy* (CE).

Dalam penelitian ini akan dilakukan pengelompokan *tweets* dari akun Twitter @LazadaID menggunakan algoritma *Modified* Gustafson-Kessel *clustering.* Penelitian akan menggunakan *tweets* yang diunggah akun Twitter @LazadaID serta jumlah *retweet* dari masing-masing *tweets* yang diunggah untuk mengetahui *tweets* yang paling disukai oleh para pengikut (*followers*) akun @LazadaID yang akan diolah menggunakan perangkat lunak RStudio dan MatLab R2015a.

2. TINJAUAN PUSTAKA

2.1. Pemasaran Media Sosial

Pemasaran media sosial adalah segala bentuk pemasaran langsung atau tidak langsung yang digunakan untuk membangun kesadaran, pengenalan, pengingatan kembali, dan pengambilan aksi terhadap sebuah merek, bisnis, produk, orang, atau hal lainnya yang dikemas menggunakan alat-alat di *social web*, seperti *blogging*, *microblogging*, *social networking*, *social bookmarking*, dan *content sharing* [7]. Banyak perusahaan yang menerapkan pemasaran media sosial dengan pertimbangan biaya (*cost*) yang cukup rendah namun memiliki pengaruh yang cukup tinggi dalam mendapatkan konsumen. Perusahaan yang menggunakan metode ini biasanya merupakan perusahaan yang bergerak di bidang *e-commerce*.

Sebelum menentukan strategi pemasaran, perusahaan harus mengetahui faktor yang mempengaruhi perilaku konsumen. Salah satu faktor yang memengaruhi perilaku pembelian *online* konsumen adalah persepsi manfaat. Pencarian *online* dan persepsi manfaat akan memberikan efek positif terhadap frekuensi belanja *online* [5]. Faktor lain yang memengaruhi adalah persepsi risiko. Persepsi risiko konsumen akan meningkat melalui dan atau besarnya hubungan konsekuensi yang negative [12]. Opini yang tersebar melalui komunitas *internet* juga akan mempengaruhi proses keputusan pembelian *online.* Sebelum pembeli memutuskan untuk melakukan pembelian barang pastinya pembeli akan mencari referensi dari orang lain sebagai bahan pertimbangan.

2.2. *Text Mining*

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 [6]. 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.

Sehingga *text mining* merupakan sebuah penemuan baru dari informasi yang belum diketahui dengan mengekstrak informasi dari sumber tertulis. Tahap-tahap *text mining* adalah sebagai berikut [6]:

a. Text Preprocessing

Text preprocessing meliputi berbagai jenis teknik ekstraksi informasi yang mengubah format mentah, tidak terstruktur, dan memiliki format asli menjadi terstruktur dan dapat diolah pada tahapan berikutnya [6]. Tahap-tahap *preprocessing* yang dilakukan antara lain:

- *Case Folding*, yaitu mengkonversi keseluruhan karakter huruf dalam dokumen menjadi huruf kecil.
- *Remove* URL, yaitu menghilangkan *link* internet (URL).
- *Remove mention*, yaitu menghilangkan rujukan kepada pengguna akun Twitter lain.
- *Unescape* HTML, yaitu menghilangkan bahasa markah yang berupa kode-kode *tag.*
- *Remove Number*, *Remove Punctuation*, dan *Remove Emoticon*, yaitu menghilangkan karakter selain huruf alphabet yang berupa angka dan tanda baca.
- Menerjemahkan kalimat dengan bahasa berbeda menjadi satu bahasa yang sama.
- *StripWhiteSpace*, yaitu menghapus spasi yang berlebih pada dokumen.
- *Tokenizing*, yaitu memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata.

b. Feature Selection

Feature Selection merupakan tahapan untuk mengurangi dimensi dari sebuah data tekstual dengan menghapus kata-kata yang tidak relevan sehingga proses pengelompokan lebih efektif dan akurat [6]. Proses yang dilakukan pada tahapan ini adalah:

- *Stopword Removal*, yaitu menghapus kata-kata yang sering muncul dalam suatu dokumen, namun memiliki arti yang tidak deskriptif dan dapat dibuang. Data *stopwords* dapat diambil dari thesis Fadillah Z Tala yang berjudul "*A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*".
- *Stemming*, yaitu proses mengubah berbagai kata berimbuhan menjadi kata dasarnya [15]. Algoritma *stemming* yang digunakan adalah algoritma *stemming* Pujangga.

c. *Text Representation*

Text representation merupakan tahapan mengubah data tekstual menjadi representasi yang lebih mudah untuk diproses. Salah satu pendekatan untuk *text representation* ini adalah dengan menggunakan matriks dokumen atau yang biasa disebut *Document Term Matrix*. Baris pada matriks mewakili dokumen yang digunakan, sedangkan kolom pada matriks berisi kata-kata, frase atau unit hasil *indexing* lainnya dalam suatu dokumen yang digunakan untuk mengetahui konteks dari dokumen tersebut (*terms*).

Untuk mempermudah tahap representasi teks dapat digunakan *wordcloud* untuk mengidentifikasi kata-kata yang ada pada data dokumen. *Word cloud* adalah presentasi grafis dari suatu dokumen, biasanya dihasilkan dengan memetakan kata-kata paling umum dari suatu dokumen dalam dua dimensi ruang, dengan frekuensi kata yang ditunjukkan oleh ukuran hurufnya [3].

Setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, sehingga perlu dilakukan pembobotan untuk setiap kata yang digunakan. TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata atau *term* terhadap suatu dokumen [13]. TF-IDF dihitung menggunakan sebagai berikut [6]:

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

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* pada dokumen ke-i.
D = banyaknya dokumen yang dibangkitkan.
d_j = banyaknya dokumen yang mengandung *term* ke j.

2.3. *Algoritma Modified Gustafson-Kessel Clustering*

Algoritma Gustafson-Kessel mengubah fungsi perhitungan jarak menjadi fungsi jarak adaptif (*adaptive distance norm*) yang selalu diperbaharui pada setiap iterasi dengan menggunakan matriks *fuzzy covariance* [8]. Algoritma Gustafson-Kessel menggunakan fungsi jarak mahalanobis sehingga lebih dapat menyesuaikan bentuk geometris untuk sebuah himpunan data, tidak seperti *Fuzzy C-Means* yang mengasumsikan bahwa bentuk geometris suatu *cluster* adalah bulat sempurna. Meskipun Gustafson-Kessel lebih unggul dari algoritma *Fuzzy C*-Means, masih terdapat masalah saat matriks *fuzzy covariance* dari data merupakan matriks *singular* maka perhitungan matriks $\boldsymbol{A}_k$ tidak dapat diterapkan [2].

Algoritma *Modified* Gustafson-Kessel *Clustering* secara lengkap adalah sebagai berikut [2]: Input data yang akan dikelompokkan sebagai X (matriks $n \times$ p), tentukan jumlah *cluster* yang akan dibentuk $(c \geq 2)$, *weighting exponent* $(m > 1)$, maksimum iterasi (t_{max}), error terkecil yang diharapkan (ε), nilai *threshold* (β), dan parameter pembobot $\gamma \in [0,1]$. Bangkitkan bilangan random $u_{ik}, 1 \leq i \leq n; 1 \leq k \leq c$ sebagai elemen-elemen matriks partisi awal $\boldsymbol{U}_0$ dan hitung matriks kovarian $\boldsymbol{F}_0$ dari keseluruhan data.

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

Ulangi untuk $t = 1, 2, \ldots, t_{max}$

***Step 1**:* Menghitung pusat *cluster* ke-k $(\boldsymbol{v}_k)$ dengan rumus:

$$\boldsymbol{v}_k^{(t)} = \frac{\sum_{i=1}^{n} \left(u_{ik}^{(t-1)}\right)^m \boldsymbol{x}_i}{\sum_{i=1}^{n} \left(u_{ik}^{(t-1)}\right)^m}, 1 \leq k \leq c \quad (3)$$

dengan:

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

***Step 2**:* Menghitung matriks kovarian *cluster* dengan rumus:

$$\boldsymbol{F_k} = \frac{\sum_{i=1}^{n} \left(u_{ik}^{(t-1)}\right)^m \left(\boldsymbol{x}_i - \boldsymbol{v}_k^{(t)}\right)\left(\boldsymbol{x}_i - \boldsymbol{v}_k^{(t)}\right)^T}{\sum_{i=1}^{n} \left(u_{ik}^{(t-1)}\right)^m}, 1 \leq k \leq c \quad (4)$$

dengan:

$\boldsymbol{x}_i$ = vektor data ke-i
$\boldsymbol{v}_k$ = pusat *cluster* ke-k.
u_{ik} = derajat keanggotaan data ke-i pada *cluster* ke-k.

m = pangkat pembobot untuk fungsi keanggotaan *fuzzy*
t = banyaknya iterasi
n = banyaknya data
c = banyaknya *cluster*

Ekstraksi nilai *eigenvectors* $\boldsymbol{\phi}$ dan *eigenvalues* λ dari $\boldsymbol{F}_k^{new}$ yang sudah dihitung dengan persamaan:

$$\boldsymbol{F}_k^{new} = (1-\gamma)\boldsymbol{F}_k + \gamma det(\boldsymbol{F}_0)^{1/p}\mathbf{I} \quad (5)$$

dengan:
γ = parameter untuk mengatur bentuk matriks *fuzzy covariance*, $\gamma \in [0,1]$
$\boldsymbol{F_0}$ = matriks kovarian dari seluruh data
$\boldsymbol{F_k}$ = matriks *fuzzy covariance cluster* ke-k (pada persamaan 4)
p = banyaknya variabel
$\boldsymbol{I}$ = matriks identitas

Jika rasio antara nilai eigen maksimal dan minimal melewati nilai *threshold* yang ditentukan, maka rekonstruksi $\boldsymbol{F_k}$ dengan penjabaran sebagai berikut:

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

dengan:
$\boldsymbol{\phi}$ = vektor eigen dari matriks *fuzzy covariance cluster* ke-k
$\boldsymbol{\Lambda}$ = matriks diagonal dari nilai-nilai eigen (*eigenvalues*) matriks *fuzzy covariance cluster* ke-k

Step 3: Menghitung jarak dengan persamaan menggunakan persaman sebagai berikut dengan $i = 1, 2, \dots, n$ dan $= 1,2, \dots, c$:

$$D^2_{ik\boldsymbol{A}_k} = \left(\boldsymbol{x}_i - \boldsymbol{v}_k^{(t)}\right)^T \left[\rho_k det(\boldsymbol{F}_k)^{1/p}\boldsymbol{F}_k^{-1}\right]\left(\boldsymbol{x}_i - \boldsymbol{v}_k^{(t)}\right) \quad (7)$$

dengan:
$D^2_{ik\boldsymbol{A}_k}$ = jarak data ke-i terhadap pusat *cluster* ke-k dengan *norm inducing matrix* $\boldsymbol{A}_k$
$\boldsymbol{x}_i$ = vektor data ke-i
$\boldsymbol{v}_k$ = pusat *cluster* ke-k.
$\boldsymbol{F}_k$ = matriks *fuzzy covarian cluster* ke-k.
ρ_k = volume *cluster* ke-k
p = banyaknya variabel
n = banyaknya data
c = banyaknya *cluster*

Step 4: Memperbaharui matriks fungsi keanggotaan

Untuk $1 \le i \le n$

Jika $D^2_{ik\boldsymbol{A}_k} > 0$ untuk $1 \le k \le c$

$$u_{ik}^{(t)} = \left[\sum_{l=1}^{c}\left(\frac{D_{ik\boldsymbol{A}_k}}{D_{il\boldsymbol{A}_k}}\right)^{\frac{2}{(m-1)}}\right]^{-1} \quad (8)$$

Jika tidak, maka:

$$u_{ik}^{(t)} = 0 \text{ jika } D^2_{ik\boldsymbol{A}_k} > 0 \text{ dan } u_{ik}^{(t)} \epsilon [0,1] \quad (9)$$
$$\text{dengan } \sum_{k=1}^{c} u_{ik}^{(t)} = 1$$

Sampai $\left\|\boldsymbol{U}^{(t)} - \boldsymbol{U}^{(t-1)}\right\| < \varepsilon$ atau jika $t >$ iterasi maksimum

Nilai *threshold* (β) yang digunakan biasanya akan ditentukan dalam angka yang besar, seperti 10^{15}. Penentuan nilai parameter pembobot γ tergantung kepada

data yang digunakan, beberapa percobaan mungkin perlu dilakukan untuk menentukan nilai γ yang tepat.

2.4. *Indeks Validitas Partition Coefficient* dan *Classification Entropy*

Validasi *cluster* mengacu kepada masalah apakah partisi yang dibentuk sudah tepat dan bagaimana mengukur ketepatan partisi. Indeks *Partition Coefficient* mengukur jumlah *overlapping* antar kelompok. Jumlah *cluster* optimal ditunjukkan oleh nilai *Partition Coefficient* yang paling besar [9]. Indeks ini didefiniskan sebagai berikut:

$$PC(c) = \frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ik})^2 \quad (10)$$

dengan:

c = banyaknya *cluster* yang dibuat.

n = banyaknya data yang dikelompokkan.

u_{ik} = derajat keanggotaan data ke-i pada *cluster* ke-k.

Indeks *Classification Entropy* (CE) digunakan untuk mengukur kekaburan (*fuzziness*) dari partisi *cluster*. Indeks ini memiliki rentang antara 0 sampai dengan $\ln(c)$. Indeks CE yang semakin kecil menunjukkan pengelompokan yang lebih baik. Indeks ini didefinisikan sebagai berikut:

$$CE(c) = \frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n}u_{ik}\,ln(u_{ik}) \quad (11)$$

dengan:

c = banyaknya *cluster* yang dibuat.

n = banyaknya data yang dikelompokkan.

u_{ik} = derajat keanggotaan data ke-i pada *cluster* ke-k.

3. METODOLOGI PENELITIAN

Pengambilan data dilakukan dengan cara *crawling* data di Twitter berupa data *tweet*, jumlah *retweet*, jumlah *like*, dan tanggal unggahan dengan bantuan Twitter API (*Application Programming Interface*). Data teks yang digunakan dalam penelitian ini adalah *tweets* dengan bahasa Indonesia dari beranda akun Twitter @LazadaID yang bukan termasuk *tweet* balasan kepada pengguna lain. *Tweets* yang digunakan merupakan *tweets* yang diunggah sebelum tanggal 2 Mei 2019 dengan jumlah *tweets* maksimal yang dapat diambil dari akun Twitter @LazadaID sebanyak 1303 *tweets*.

Analisis pengelompokan *tweets* dilakukan menggunakan metode *text mining* dengan metode lanjutan berupa Algoritma *Modified* Gustafson-Kessel. Data penelitian diproses menggunakan perangkat lunak *RStudio* dan MatLab R2015a. Perangkat lunak *RStudio* digunakan dalam proses *crawling* data hingga terbentuk *document term matrix*. Sedangkan perangkat lunak MatLab R2015a digunakan untuk proses *clustering* data. Setelah didapatkan *cluster tweets* berdasarkan konten selanjutnya dilakukan *profiling* masing masing *cluster* dengan bantuan *wordcloud*.

4. HASIL DAN PEMBAHASAN

Akun Twitter @LazadaID pertama kali diluncurkan pada bulan Maret 2012 dengan jumlah *tweets* sebanyak 127 ribu *tweet* serta jumlah pengikut sebanyak 366 ribu akun per Mei 2019, menjadikan akun Twitter @LazadaID sebagai perusahaan *e-commerce* dengan jumlah pengikut terbanyak di Indonesia.

Pada saat bergabung dengan Twitter API untuk proses *crawling* data, akan didapatkan beberapa kode berupa *consumer key*, *consumer secret*, *access token*, dan *access key* yang digunakan sebagai hak akses *stream* untuk mengambil data *text* Twitter dengan

software R. Pengambilan data dilakukan menggunakan fungsi '`tweets<-userTimeline('LazadaID',n=3200,excludeReplies = TRUE)`'.

Tabel 1. Contoh Tweet dari Beranda Twitter @LazadaID

No	*Text*	Tanggal *Tweet*	Jumlah *Retweet*	Jumlah *Like*
1	Hi Lazadians! Jangan lupa rawat kulitmu agar nutrisinya terpenuhi ya! Cek disini >> https://t.co/iTDVmxB8hO	5/2/2019 7:34 AM	0	2
2	Ada penawaran menarik nih dari @Xiaomi! Yuk cek sekarang https://t.co/vWzNAGa9Mx	5/2/2019 7:29 AM	0	4

4.1. Pengolahan Data Teks Menjadi *Document Term Matrix*

Data teks diolah melalui tahapan *text pre-procesing*, *feature selection*, dan *text representation* untuk merubah bentuk teks menjadi suatu matriks angka yang digunakan saat proses *clustering*.

a. Text Preprocessing

Text preprocessing mengubah format seluruh *tweets* menjadi suatu data tekstual yang memiliki format sama. Pada tahap ini *link* internet, *mention*, HTML, kata yang diawali dengan "laz", dan karakter selain huruf alphabet yang ada pada seluruh *tweets* akan dihapus. Selanjutnya, karena ada beberapa *tweets* yang menggunakan bahasa Inggris, maka perlu dilakukan proses menerjemahkan kata ke dalam bahasa Indonesia.

Tabel 2. Contoh Hasil Proses Text Preprocessing

No. *Tweets*	*Tweets* Hasil *Text Preprocessing*
1	hi jangan lupa rawat kulitmu agar nutrisinya terpenuhi ya cek disini
2	ada penawaran menarik nih dari xiaomi yakin mau yuk cek sekarang

b. Feature Selection

Feature selection terdiri dari 2 proses yaitu *stopword removal* dan *stemming*. Pada proses *stopword removal*, penghapusan kata-kata yang dianggap tidak penting atau tidak bermakna yang sudah *diinput* ke dalam kamus *stoplist* dilakukan dengan fungsi '`data <- tm_map(data, removeWords, cStopwordID)`'. Untuk proses *stemming* dilakukan dengan menggunakan algoritma Pujangga dengan bantuan *interface* InaNLP (*Indonesian Natural Language Processing*).

Tabel 3. Contoh Hasil Proses Feature Selection

No. *Tweets*	*Tweets* Hasil *Feature Selection*
1	lupa rawat kulit nutrisi penuhi
2	penawaran menarik xiaomi

c. Text Representation

Pada proses ini, dilakukan perubahan data *tweet* menjadi matriks yang berisi frekuensi kemunculan kata (TF) pada sebuah dokumen, serta pembobotannya dengan menggunakan pembobotan TF-IDF. *Document Term Matrix* dengan pembobotan TF digunakan untuk mellihat kecenderungan *term* yang sering muncul pada 1303 *tweets* dari akun Twitter @LazadaID. Sedangkan *Document Term Matrix* dengan pembobotan TF-IDF digunakan untuk melakukan *clustering* dengan algoritma *Modified* Gustafson-Kessel.

Berdasarkan hasil *text representation*, jumlah kata yang menyusun 1303 *tweets* dari akun Lazada Indonesia adalah 1364 kata. Seluruh kata tersebut akan menjadi variabel dari tiap *tweet*, dengan komponen dari matriks berupa jumlah dari suatu kata pada tiap

tweet.Proses perubahan tersebut dilakukan dengan menggunakan fungsi 'as.matrix(weightTfIdf(m= DocumentTermMatrix(data), normalize = TRUE))'.

Tabel 4. Hasil Document Term Matrix Pembobotan TF

No	*Tweet*	bayi	belanja	diskon	...	promo	unilever	voucher
95	maybelline diskon voucher tambah belanja	0	1	1	...	0	0	1
717	unilever voucher diskon langsung unilever diskon	0	0	2	...	0	2	1

Tabel 5. Hasil Document Term Matrix Pembobotan TF-IDF

No	*Tweet*	bayi	belanja	diskon	...	promo	unilever	voucher
95	maybelline diskon voucher tambah belanja	0	0,603	0,461	...	0	0	0,632
717	unilever voucher diskon langsung unilever diskon	0	0	0,658	...	0	1,682	0,451

4.2. Algoritma *Modified* Gustafson-Kessel *Clustering*

Proses *clustering* dengan algoritma *Modified* Gustafson-Kessel *clustering* dilakukan dengan menggunakan fungsi 'result = MGK(Z,U0,m,tol,beta,gamma)' pada *command window* MatLab 2015a. Data yang digunakan adalah *Document Term Matrix* dengan pembobotan TF-IDF. Pangkat *fuzzyfier* yang digunakan adalah $m = 3{,}75$, nilai tersebut merupakan hasil yang paling optimal dari proses *trial and error* karena menghasilkan fungsi keanggotaan $u_{ik} \neq {}^{1}/_{c}$. Untuk batas *error* terkecil dan nilai *threshold* yang digunakan sebesar $\varepsilon = 10^{-3}$ dan $\beta = 10^{15}$ [2]. Sedangkan nilai parameter pembobot yang digunakan sebesar $\gamma = 0{,}3$. Proses c*lustering* dilakukan sebanyak 5 kali dengan jumlah *cluster* yang berbeda-beda, dimulai dari jumlah *cluster* sebanyak 2 sampai dengan 6 *cluster*.

Jumlah *cluster* optimum dari proses *clustering* ditentukan menggunakan indeks validitas *Partition Coefficient* (PC) dan *Classification Entropy* (CE). Hasil yang didapat yaitu untuk jumlah *cluster* sebanyak 4, 5, dan 6 *cluster*, hanya 3 *cluster* dari total seluruh *cluster* yang memiliki anggota kelompok. Jadi, dapat disimpulkan bahwa *tweets* dari akun @LazadaID hanya dapat dikelompokkan dengan jumlah *cluster* maksimal sebanyak 3 *cluster*. Sehingga dalam penelitian ini akan dilakukan pengukuran validitas *Partition Coefficient* (PC) dan *Classification Entropy* (CE) terhadap pengelompokan dengan jumlah *cluster* 2 dan 3 saja.

Tabel 6. Nilai Partition Coefficient dan Classification Entropy

Jumlah *Cluster*	*Partition Coefficient* (PC)	*Classification Entropy* (CE)
2	0,048744	1,098612
3	0,148658	0,693143

Berdasarkan Tabel 6 dapat dilihat bahwa jumlah *cluster* optimum diberikan untuk jumlah *cluster* sebanyak 3 *cluster* karena memiliki nilai *Partition Coefficient* (PC) yang paling besar dan nilai *Classification Entropy* (CE) paling kecil dibandingkan dengan nilai pada jumlah *cluster* sebesar 2. Jadi, jumlah *cluster* yang akan dianalisis pada penelitian ini adalah sebanyak 3 *cluster*. Hasil *clustering* dengan jumlah 3 *cluster* dapat dilihat pada Tabel 7.

Tabel 7. Hasil Clustering dengan Algoritma Gustafson-Kessel

Cluster ke-	Nomor Anggota *Tweet*	Jumlah Anggota

1	19, 135, 137, 183, 191, 201, 206, 234, 269, 408, 462, 487, 515, 777, 969, 988, 1023, 1178, 1226, 1242, 1246, 1252	22
2	2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 18, 21, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 40, 41, 42, 45, …, 1226, 1242, 1246, 1252	787
3	1, 6, 8, 12, 14, 16, 17, 20, 22, 23, 24, 29, 32, 38, 39, 43, 44, 47, 48, 53, 59, 66, 68, 69, 70, …, 1294, 1298, 1300, 1303	494

Analisis lanjut untuk menginterpretasikan setiap *cluster* pada penelitian ini dibantu dengan *wordcloud* seperti pada Gambar 1. *Wordcloud* digunakan untuk melihat kata-kata yang menyusun masing-masing *cluster*. Dari *wordcloud* pada Gambar 1 terlihat bahwa pada *cluster* 1 kata "iphone" memiliki ukuran paling besar, sehingga dapat disimpulkan bahwa kata "iphone" memiliki frekuensi kemunculan paling tinggi. Setelah ditinjau kembali dengan mencari *tweet* yang berisi kata-kata berwarna kuning dan biru dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 1 adalah *tweet* yang berisi penawaran barang-barang elektronik terbaru yang tersedia di Lazada seperti iphone, kabel, *earphone*, aukey, *powerbank*, *charger*, vivo, dan sebagainya.

Gambar 1. Wordcloud Cluster 1

Dari *wordcloud* pada Gambar 2 terlihat bahwa pada *cluster* 2 kata yang paling banyak muncul adalah "diskon", "promo", "produk", "belanja", dan "*voucher*". Kata yang berwarna biru merupakan kata dengan frekuensi kemunculan yang lebih rendah dibandingkan kata dengan warna lainnya. Setelah ditinjau kembali dengan mencari *tweet* dengan kata-kata yang sering muncul dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 2 adalah *tweet* yang berisi penawaran mengenai diskon dan promo di Lazada. Kata diskon di dalam *cluster* 2 muncul sebanyak 220 kali dalam 202 *tweets*, sedangkan untuk kata promo muncul sebanyak 177 kali dalam 169 *tweets*. Untuk *tweets* yang berisi kata "belanja" hampir semuanya merujuk kepada ajakan untuk berbelanja di Lazada karena sedang diadakan banyak promo dan diskon.

Gambar 2. Wordcloud Cluster 2

Dari *wordcloud* pada Gambar 3 terlihat bahwa pada *cluster* 3 kata yang paling banyak muncul adalah "selamat", "menang", "*voucher*", "belanja", dan "hadiah". Kata yang berwarna biru merupakan kata dengan frekuensi kemunculan yang lebih rendah dibandingkan kata dengan warna lainnya. Setelah ditinjau kembali dengan mencari *tweet* dengan kata-kata yang sering muncul dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 3 adalah *tweet* yang berisi tentang kuis berhadiah. Kata "selamat" di dalam *cluster* 3 muncul sebanyak 118 kali dalam 114 *tweets*, sedangkan untuk kata "menang" muncul sebanyak 110 kali dalam 109 *tweets*.

Gambar 3. Wordcloud Cluster 3

4.3. Penentuan Konten yang Disukai *Followers* @LazadaID

Konten yang disukai oleh para *followers* akun Twitter @LazadaID dapat ditentukan berdasarkan perhitungan rata-rata jumlah *retweet* dan *like* masing-masing *cluster*. Hal tersebut dikarenakan jika pengguna Twitter menyukai suatu *tweet* maka kemungkinan yang akan dilakukan pengguna tersebut adalah melakukan *retweet* atau menekan tombol *like*.

Tabel 8. . Hasil Perhitungan Rata-rata Jumlah Retweet dan Like

Nomor *Cluster*	Jumlah *Tweet*	Jumlah *Retweet*	Jumlah *Like*	Rata-rata *Retweet*	Rata-rata *Like*
1	22	570	566	25,90909	25,72727
2	787	26387	20664	33,52859	26,25667
3	494	21182	15664	42,87854	31,7085

Pada Tabel 8 terlihat bahwa *cluster* yang memiliki rata-rata *retweet* dan *like* tertinggi adalah *cluster* 3 yaitu *tweet* mengenai kuis berhadiah, sedangkan jenis konten yang memiliki rata-rata *retweet* dan *like* terendah adalah *cluster* 1 yaitu *tweet* mengenai penawaran barang-barang elektronik. Berdasarkan hasil yang didapatkan terlihat bahwa konsumen dari perusahaan *e-commerce* Lazada Indonesia lebih tertarik terhadap program kuis berhadiah dibandingkan dengan diskon ataupun penawaran barang-barang elektronik. Oleh karena itu, PT Lazada Indonesia diharapkan mampu memanfaatkan konten kuis berhadiah sebagai sarana *advertising* pada platform Twitter semaksimal mungkin. Kuis berhadiah mampu memberikan dorongan kepada konsumen untuk melakukan pembelian dalam jumlah yang lebih banyak dan dapat meminimumkan perilaku berganti-ganti merek [4].

5. KESIMPULAN

Penerapan algoritma Gustafson-Kessel untuk *clustering tweets* pada akun Twitter @LazadaID menghasilkan 3 *cluster tweet* dengan konten mengenai penawaran barang elektronik, diskon dan promo, serta mengenai kuis berhadiah. Jumlah 3 *cluster tweet* yang dibentuk ditentukan berdasarkan nilai indeks validitas *Partition Coefficient* (PC) terbesar dan nilai *Classification Entropy* (CE) terkecil. Dari ketiga *cluster* tersebut didapatkan jenis

konten dengan rata-rata jumlah *retweet* dan *like* tertinggi yaitu *tweets* mengenai kuis berhadiah, serta rata-rata terendah yaitu mengenai penawaran barang elektronik. Oleh karena itu, PT Lazada Indonesia dapat menggunakan *tweets* dengan konten kuis berhadiah sebagai sarana *advertising* pada *platform* media sosial Twitter karena *tweet* tersebut banyak disukai oleh para *followers* @LazadaID.

Ada beberapa perbaikan yang dapat dilakukan pada penelitian selanjutnya, yaitu menggunakan metode *clustering* yang lebih efisien serta memiliki kecocokan dengan bentuk data yang relatif homogen. Kemudian untuk menambah akurasi dari tahapan *preprocessing* data dapat digunakan *composite tokenization* dalam proses *tokenizing*.

DAFTAR PUSTAKA

[1] Aseanup, 2019. *Top 10 E-commerce Sites in Indonesia 2019*. https://aseanup.com/top-e-commerce-sites-indonesia/. Diakses 1 Maret 2019

[2] Babuska, R., Veen, P. v. d. & Kaymak, U., 2002. *Improved Covariance Estimation for Gustafson-Kessel Clustering.* Netherlands: Delft University of Technology.

[3] Castella, Q. & Sutton, C., 2014. *Word Storms: Multiples of Word Clouds for Visual Comparison of Documents*. Seoul, International Conference on World Wide Web, Vol. 1.

[4] F. Tjiptono, *Brand Management & Strategy*. Yogyakarta: Andi Offset, 2005.

[5] Farag, S. & Lyons, G. D., 2007. *Conceptualising barriers to travel information use.* United Kingdom, Proceedings of the 39th Annual Universities.

[6] Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook.* New York: Cambridge University Press.

[7] Gunelius, S., 2011. *30 Minute Social Media Marketing.* United States: McGraw Hill.

[8] Gustafson, D. & Kessel, W., 1979. *Fuzzy Clustering with a Fuzzy Covariance Matrix.* San Diego, Hal. 761-766.

[9] Jansen, S., 2007. *Customer Segmentation and Costumer Profiling for a Mobile Telecommunications Company Based on Usage Behavior :A Vodafone Case Study.* Maastricth: University of Maastricth.

[10] Kemenkominfo, 2019. *Kominfo : Pengguna Internet di Indonesia 63 juta.* https://www.kominfo.go.id/content/detail/3415/kominfo-pengguna-internet-di-indonesia-63-juta-orang/0/berita_satker. Diakses 1 Maret 2019.

[11] Littlejohn, S. W. & Foss, K. A., 2009. *Teori Komunikasi.* Jakarta: Salemba Humanika.

[12] Oglethorpe, J. E. & Monroe, K. B., 2008. *Determinants of Perceived Health and Safety Risks of Selected Hazardous Products and Activities*. The Journal of Consumer Affair, Vol. 28(2), Hal. 326-346.

[13] Robertson, S., 2005. *Understanding inverse document frequency: On theoretical arguments for IDF*. Journal of Documentation, Hal. 502-520.

[14] Santosa, B., Conway, T. & Trafalis, T., 2007. *A Hybrid Knowledge Based Clustering Multiclass SVM Approach for Genes Expression Analysis.*. Boston: Springer.

[15] Tala, F. Z., 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Netherland: Institute for Logic, Language, and Computation Universiteit van Amsterdam.