ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 1, Tahun 2022, Halaman 86 - 98
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

IMPLEMENTASI ALGORITMA *FUZZY C-MEANS* DAN *FUZZY POSSIBILISTICS C-MEANS* UNTUK KLASTERISASI DATA *TWEETS* PADA AKUN TWITTER TOKOPEDIA

Ghina Nabila Saputro Putri[1], Dwi Ispriyanti[2*], Tatik Widiharih[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***email:** dwiispriyanti@yahoo.com

ABSTRACT

Social media has become the most popular media, which can be accessed by young to old age. Twitter became one of the effective media and the familiar one used by the public, thus making the company make Twitter one of the promotional tools, one of which is Tokopedia. The research aims to group tweets uploaded by @tokopedia Twitter accounts based on the type of tweets content that gets a lot of retweets and likes by followers of @tokopedia. Application of text mining to cluster tweets on the @tokopedia Twitter account using Fuzzy C-Means and Fuzzy Possibilistic C-Means algorithms that viewed the accuracy comparison of both methods used the Modified Partition Coefficient (MPC) cluster validity. The clustering process was carried out five times by the number of clusters ranging from 3 to 7 clusters. The results of the study showed the Fuzzy C-Means method is a better method compared to the Fuzzy Possibilistic C-Means method in clustering data tweets, with the number of clusters formed is 4. The content type formed is related to promo, discount, cashback, prize quizzes, and event promotions organized by Tokopedia. Content with the highest average number of retweets and likes is about automotive deals, sports tools, and merchandise offerings. So, that PT Tokopedia can use this content type as a tool for advertising on Twitter because it gets more likes by followers of @tokopedia.

Keywords: Data *Tweets*, *Clustering*, *Fuzzy C-Means*, *Fuzzy Possibilistics C-Means*, *Modified Partition Coefficient*.

1. PENDAHULUAN

Pada era globalisasi dan perkembangan teknologi kearah serba digital saat ini semakin pesat. Media sosial *online* sudah menjadi media yang paling populer, yang dapat diakses oleh berbagai kalangan dari usia muda sampai usia tua. Direktur Pelayanan Informasi Internasional Ditjen Informasi dan Komunikasi Publik (IKP) mengatakan, situs jejaring sosial yang paling banyak diakses adalah Facebook dan Twitter. Twitter merupakan salah satu media yang efektif dan juga merupakan salah satu media sosial yang akrab digunakan oleh masyarakat Indonesia. Berangkat dari hal tersebut banyak perusahaan melihat itu sebagai salah satu peluang, yang akhirnya menggunakan Twitter sebagai alat pemasaran dan media untuk menyebarkan informasi. Salah satu perusahaan yang menggunakan Twitter sebagai alat promosi adalah PT Tokopedia dengan nama akun Twitter @tokopedia. Perusahaan tersebut dapat melihat konten yang disukai oleh para *followers*-nya dengan melihat jumlah *retweet* dan *like* pada setiap *tweets* yang diunggah pada akun Twitter tersebut.

Clustering merupakan proses pembagian data ke dalam kelas atau *cluster* berdasarkan tingkat kesamaannya. Data-data yang memiliki kemiripan karakteristik akan berkumpul dalam kelompok atau *cluster* yang sama. Sementara data-data yang memiliki perbedaan karakteristik, akan berkumpul dalam kelompok atau *cluster* yang berbeda (Siyamto, 2017). Pada analisis *cluster* mengukur kemiripan karakteristik antar objek yang diteliti dapat menggunakan ukuran jarak. Semakin kecil jarak yang diperoleh, maka semakin dekat letak objek dengan pusat *cluster*. Hal lain yang perlu diperhatikan pada proses *clustering* adalah

penentuan banyaknya *cluster*. Penentuan ini dapat berpengaruh pada tingkat validitas hasil *clustering*. Validitas tersebut dilakukan agar memperoleh jumlah *cluster* yang sesuai dengan data yang sedang digunakan.

Pada penelitian ini akan mengkaji perbandingan hasil *clustering* dengan metode *Fuzzy C-Means* dan *Fuzzy Possibilistics C-Means*. Metode *Fuzzy C-Means* adalah suatu teknik *clustering* data dimana keberadaan tiap tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaan (Kusumadewi & Purnomo, 2013), sedangkan *Fuzzy Possibilistic C-Means* (FPCM) juga menghasilkan nilai derajat keanggotaan dan nilai kekhasan untuk menentukan setiap titik data yang termasuk pada *cluster* tertentu. Penelitian ini akan mengelompokkan *tweets* dari akun Twitter @tokopedia serta melihat jumlah *retweet* dan *like* dari masing-masing *cluster* untuk mengetahui *tweets* yang paling disukai oleh para pengikut akun @tokopedia. Keakuratan metode ini akan dibandingkan dengan indeks validitas *Modified Partition Coefficient* (MPC) sehingga akan didapatkan metode yang tepat dalam pengelompokkan data *tweets*.

2. TINJAUAN PUSTAKA

2.1. PT Tokopedia

Tokopedia adalah perusahaan teknologi Indonesia dengan misi untuk demokratisasi perdagangan melalui teknologi dan mendorong jutaan pedagang dan konsumen untuk berpartisipasi dalam masa depan perdagangan. Tokopedia mempunyai visi yaitu membangun ekosistem di mana semua orang dapat memulai dan menemukan apapun dengan mudah (Tokopedia, 2021).

2.2. Pemasaran Melalui Media Sosial

Bagi setiap individu, motivasi menggunakan media sosial adalah mencari informasi, berbagi informasi, hiburan, relaksasi, dan interaksi sosial (Whiting & Williams, 2013). Bagi organisasi atau perusahaan, media sosial banyak digunakan sebagai media atau alat untuk melakukan komunikasi pemasaran. Saat ini sudah banyak media sosial yang dapat digunakan untuk pemasaran atau promosi produk atau jasa. Media sosial yang sering digunakan untuk pemasaran atau promosi di antaranya Facebook, Instagram, dan Twitter. Melalui media macam ini, perusahaan atau *brand* dapat melakukan promosi yang terstruktur dan tepat sasaran.

2.3. Twitter

Twitter adalah layanan jejaring sosial online dan layanan *microblogging* yang memungkinkan penggunanya untuk mengikuti aktivitas pengguna lain, membaca, dan memposting *tweet* (Twitter Inc., 2021). Twitter memberikan akses kepada penggunanya untuk mengirimkan sebuah pesan singkat yang terdiri dari maksimal 280 karakter yang dikenal dengan sebutan *tweet*. *Tweet* sendiri bisa terdiri dari pesan teks dan foto.

2.4. Twitter *Crawling*

Menurut Liu (2011) dalam (Hanifah & Nurhasanah, 2018) *Crawling* adalah proses menjelajahi web dan mengunduh halaman web secara otomatis untuk mengumpulkan informasi. Pengumpulan data yang diunduh dari server Twitter berupa *user* dan *tweet* beserta atribut-atributnya. API Twitter atau *Application Programming Interface* (API) adalah suatu program atau aplikasi yang disediakan oleh Twitter untuk mempermudah developer lain dalam mengakses informasi yang ada di website Twitter. Pengembang dapat mengakses

tweet dengan mencari kata kunci tertentu, atau meminta sampel tweet dari akun tertentu (Twitter, 2020).

2.5. *Text Preprocessing*

Text Preprocessing adalah tahap awal dari sistem *text mining* yang meliputi berbagai jenis teknik yang diadaptasi dari pengambilan informasi dan ekstraksi informasi yang mengubah format mentah, tidak terstruktur, dan memiliki format asli menjadi terstruktur (Feldman & Sanger, 2007). Tahapan dalam pre-processing yaitu sebagai berikut: *case folding*, remove url, *unescape HTML, remove mention, remove number, remove punctuation, remove emoticon, strip white space,* dan normalisasi kata.

2.6. *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 (Feldman & Sanger, 2007). Tahapan dalam feature selection yaitu, stopwords removal, stemming, dan tokenizing. Tahap stopwors removal ialah menghilangkan kata-kata umum yang tidak berhubungan dengan subyek utama, seperti kata depan dan konjungsi. Tahap stemming ialah mengubah berbagai kata imbuhan menjadi kta dasarnya (Tala, 2003). Sedangkan, tahap Tokenizing adalah pemecahan kalimat menjadi potongan kata-kata.

2.7. *Text Representation*

Text representation merupakan tahapan merubah data tekstual menjadi representasi yang lebih mudah untuk diproses. Data tekstual yang tidak terstruktur akan direpresentasikan secara numerik untuk membuatnya dapat dihitung secara matematis (Yan, 2009). Salah satu pendekatan untuk *text representation* ini adalah dengan menggunakan matriks dokumen atau yang biasa disebut *Document Term Matrix*. *Document Term Matrix* ialah menunjukkan hubungan antara *terms* dan dokumen, dimana setiap baris mewakilkan 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*).

Hal yang paling utama dalam *text mining* ialah bagaimana mengukur tentang apa dokumen itu, maka perlu dilakukan pembobotan untuk setiap kata yang digunakan. Metode TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu *term* terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut (Nurjannah, et al., 2013). TF-IDF dihitung menggunakan rumus seperti berikut:

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

$W_{j,i}$ adalah TF-IDF, $n_{j,i}$ adalah jumlah kemunculan *term* j dalam dokumen ke i, $\sum_k n_{k,i}$ adalah jumlah kemunculan seluruh *term* pada dokumen ke i, D adalah Banyaknya dokumen yang digunakan, dan d_j adalah banyaknya dokumen yang mengandung *term* j.

2.8. *Word Cloud*

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 (Castella dan Sutton, 2014). Menurut McNaught & Lam (2010) visualisasi *word cloud* dapat memudahkan pengamat memiliki gambaran umum tentang topik utama dan tema utama dalam sebuah teks, dan dapat menggambarkan sudut pandang utama yang dimiliki oleh penulis teks sehingga dapat menjadi alat bantu dalam melakukan analisis terhadap sebuah wacana tertulis.

2.9. *Fuzzy C-Means*

Fuzzy C-Means adalah suatu teknik *clustering* data dimana keberadaan tiap-tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaan (Kusumadewi & Purnomo, 2013). Konsep dasar FCM, pertama kali adalah menentukan pusat *cluster*, yang akan menandai lokasi rata-rata untuk tiap-tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimisasi fungsi objektif yang menggambarkan jarak dari titik data yang akan diberikan ke pusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut.

Adapun algoritma *Fuzzy C-Means* (FCM) adalah sebagai berikut:

1. Memasukkan data (X_{ij}) yang akan di*cluster* ke dalam sebuah matriks, dimana matriks berukuran $s \times p$, dengan *s* adalah banyaknya data yang akan di*cluster* dan *p* adalah banyaknya atribut atau variabel.
2. Menentukan jumlah *cluster* (c), pangkat (w), maksimum iterasi, *error* terkecil yang diharapkan (ξ), fungsi objektif awal (P_0), dan iterasi awal (*t*).
3. Membangkitkan bilangan random a_{ik}, antara 0-1 sebagai elemen matriks keanggotaan awal U.

$$\mathrm{U}^{(0)} = \begin{bmatrix} a_{11} & \cdots & a_{1c} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{sc} \end{bmatrix}$$

Hitung jumlah setiap baris:

$$Q_i = \sum_{k=1}^{c} a_{ik}$$

dengan *i* = 1,2,…,s.

Hitung:

$$\mu_{ik}{}^{(0)} = \frac{a_{ik}}{Q_i} \qquad (2)$$

4. Menghitung pusat *cluster* ke-*k*. Pusat *cluster* dilambangkan dengan V_{kj}.

$$V_{kj} = \frac{\sum_{i=1}^{s}\left((\mu_{ik})^{w} X_{ij}\right)}{\sum_{i=1}^{s}(\mu_{ik})^{w}} \qquad (3)$$

5. Menghitung fungsi objektif pada iterasi ke-*t*

$$P_t = \sum_{i=1}^{s}\sum_{k=i}^{c}\left(\left[\sum_{j=1}^{p} {d_{ik}}^2(x_{ij}, v_{kj})^2\right](\mu_{ik})^w\right) \quad (4)$$

6. Menghitung perubahan matriks keanggotaan:

$$\mu_{ik}{}^{(t)} = \frac{\left[\sum_{j=1}^{p} {d_{ik}}^2(x_{ij}, v_{kj})\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{p} {d_{ik}}^2(x_{ij}, v_{kj})\right]^{\frac{-1}{w-1}}} \quad (5)$$

7. Mengecek kondisi berhenti:
   Jika $|P_t - P_{t-1}| < \xi$ atau $t > MaxIter$ maka berhenti;
   Jika tidak: $t + 1$, ulangi langkah ke-4
8. Menentukan anggota pada setiap *cluster*
   Objek data dikatakan masuk dalam suatu *cluster* jika nilai kenaggotaan mendekati 1 atau yang paling besar.

2.10. *Fuzzy Possibilistics C-Means*

Fuzzy Possibilistic C-Means (FPCM) diperkenalkan pertama kali oleh Pal dkk (1997) yaitu menggunakan kelebihan dari pemodelan *fuzzy* dan *possibilistic*, yang dengan demikian dapat mengurangi kelemahan dari keduanya. Untuk memperbaiki sub sturuktur data, metode ini mengadopsi dua jenis keanggotaan yaitu kekhasan relatif (*fuzzy*) dan kekhasan absolut (*possibilistic*). Adapun matriks keanggotaan absolut ini dapat mengurangi dampak dari *outlier* (Chaudhuri, 2015).

Adapun algoritma dalam *Fuzzy Possibilistic C-Means* (FPCM) adalah:

1. Memasukkan data (X_{ij}) yang akan di*cluster* ke dalam sebuah matriks, dimana matriks berukuran $s \times p$, dengan *s* adalah banyaknya data yang akan di*cluster* dan *p* adalah banyaknya atribut atau variabel.
2. Menentukan jumlah *cluster* (c), pangkat (w), maksimum iterasi, error terkecil yang diharapkan (ξ), fungsi objektif awal (P_0), dan iterasi awal (*t*).
3. Memanggil hasil akhir yang berupa matriks keangotaan dan pusat *cluster* (V_{kj}) pada algoritma FCM, untuk menghitung matriks kekhasan absolut, T, sebagai berikut:

$$\mathrm{T} = \begin{bmatrix} t_{11} & \cdots & t_{1c} \\ \vdots & \ddots & \vdots \\ t_{s1} & \cdots & t_{sc} \end{bmatrix}$$

4. Memperbaiki pusat *cluster* (V_{kj}) dengan perhitungan sebagai berikut:

$$V_{kj} = \frac{\sum_{i=1}^{s}(\mu_{ik}{}^w + t_{ik}{}^\eta)\, X_{ij}}{\sum_{i=1}^{s}(\mu_{ik}{}^w + t_{ik}{}^\eta)} \quad (6)$$

5. Menghitung fungsi objektif pada iterasi ke-*l*

$$P_l = \sum_{i=1}^{s}\sum_{k=i}^{c}\left(\left[\sum_{j=1}^{p} {d_{ik}}^2(x_{ij}, v_{kj})^2\right](\mu_{ik}{}^w + t_{ik}{}^\eta)\right) \quad (7)$$

6. Menghitung perubahan matriks keanggotaan pada Persamaan (9) dengan perhitungan sebagai berikut:

$$\mu_{ik}{}^{(l)} = \frac{\left[\sum_{j=1}^{p} d_{ik}{}^{2}(x_{ij}, v_{kj})^{2}\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{p} d_{ik}{}^{2}(x_{ij}, v_{kj})^{2}\right]^{\frac{-1}{w-1}}} \quad (8)$$

7. Menghitung perubahan matriks kekhasan absolut pada Persamaan (11) dengan perhitungan sebagai berikut:

$$t_{ik}{}^{(l)} = \frac{\left[\sum_{j=1}^{p} d_{ik}{}^{2}(x_{ij}, v_{kj})^{2}\right]^{\frac{-1}{\eta-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{p} d_{ik}{}^{2}(x_{ij}, v_{kj})^{2}\right]^{\frac{-1}{\eta-1}}} \quad (9)$$

8. Mengecek kondisi berhenti:
   Jika $|P_l - P_{l-1}| < \xi$ atau $l > MaxIter$ maka berhenti;
   Jika tidak: $l + 1$, ulangi langkah ke-4
9. Menentukan anggota pada setiap *cluster*:
   Objek data dikatakan masuk dalam suatu *cluster* jika nilai kenaggotaan mendekati 1 atau yang paling besar.

2.11. Validitas *Cluster*

Indeks validasi yang hanya dengan menghitung jarak antara derajat keanggotaan dan pusat *cluster* ialah indeks *Modified Partition Coefficient* (MPC). MPC merupakan modifikasi dari indeks *Partition Coefficient* (PC) yang mampu mengurangi perubahan yang monoton pada PC. Sehingga dengan diketahuinya indeks MPC maka dapat memvalidasi jumlah *cluster* yang tepat (Suleman, 2015).

Metode *Modified Partition Coefficient* (MPC) didefinisikan dengan persamaan:

$$MPC\ (c) = 1 - \frac{c}{c-1}(1 - PC(c)) \quad (10)$$

Adapun persamaan nilai indeks PC ini adalah:

$$PC\ (c) = \frac{1}{s}\sum_{k=1}^{c}\sum_{i=1}^{s}\mu_{ik}{}^{2} \quad (11)$$

3. METODE PENELITIAN

3.1. Jenis dan Sumber Data

Data yang digunakan pada penelitian ini adalah data primer yang merupakan data kualitatif, yaitu data hasil *Twitter Crawling* berupa data *tweets*, jumlah *retweet*, dan jumlah *like* pada akun Twitter @Tokopedia. *Twitter Crawling* dilakukan pada tanggal 22 Februari 2021 dengan *tweets* yang berhasil ter-*crawling* sejumlah 1164 *tweets.*

3.2. Variabel Peneltian

Variabel penelitian yang digunakan dalam penelitian ini adalah *tweets* yang diunggah oleh akun Twitter @tokopedia yang di-*crawling* pada tanggal 22 Februari 2021.

3.3. Tahapan Pengolahan Data

Langkah-langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. *Twitter Crawling*

2. *Data Pre-Processing* (*case folding*, *remove* URL, *unescaped* HTML, *remove mention*, *remove number*, *remove punctuation*, *remove emoticon*, *strip white space*, dan normalisasi kata)
3. *Feature selection* (*Stopwords removal*, *Stemming*, dan *Tokenizing*)
4. Pembobotan kata dengan TF-IDF
5. Tahap *Clustering*
   a. Algoritma *Fuzzy C-Means*
      1. Menentukan jumlah *cluster* yang akan dibentuk, pangkat, maksimum iterasi, *error*, fungsi objektif awal, dan iterasi awal.
      2. Membangkitkan bilangan *random* pada matriks keanggotaan awal U
      3. Menghitung pusat *cluster* (V_{kj})
      4. Menghitung fungsi objektif pada iterasi ke-t (P_t)
      5. Menghitung perubahan matriks keanggotaan
      6. Mengecek kondisi berhenti:
         Jika $|P_t - P_{t-1}| < \xi$ atau $t > MaxIter$ maka berhenti;
         Jika tidak: $t + 1$, ulangi menghitung pusat *cluster*
      7. Menentukan anggota pada setiap *cluster*
   b. Algoritma *Fuzzy Possibilistics C-Means*
      1. Memanggil hasil akhir matriks keanggotaan (μ_{ik}) dan pusat *cluster* (V_{kj}) pada algoritma FCM untuk menghitung matriks kekhasan absolut (t_{ik})
      2. Memperbaiki pusat *cluster* (V_{kj})
      3. Menghitung fungsi objektif pada iterasi ke-l (P_l)
      4. Menghitung perubahan matriks keanggotaan
      5. Menghitung perubahan matriks kekhasan absolut
      6. Mengecek kondisi berhenti
         Jika $|P_l - P_{l-1}| < \xi$ atau $l > MaxIter$ maka berhenti;
         Jika tidak: $l + 1$, ulangi menghitung pusat *cluster*
      7. Menentukan anggota pada setiap *cluster*
6. Perbandingan metode *Fuzzy C-Means* dan *Fuzzy Possibilistics C-Means*.
7. Menentukan *cluster* optimum dengan metode *Modified Partition Coefficient* dengan memilih nilai MPC yang paling besar.
8. Membuat *profiling* dari setiap *cluster* menggunakan *word cloud*
9. Intepretasi hasil *profiling* dari setiap *cluster*

4. HASIL DAN PEMBAHASAN

4.1. Twitter *Crawling*

Pengambilan data dilakukan secara *real time* dari Twitter dengan ketentuan *tweet* yang diambil adalah *tweets* dari beranda akun @tokopedia. Tweet yang diambil secara otomatis akan tersimpan dalam bentuk list. Tweet yang berhasil terambil berjumlah 1164 *tweets* dan disimpan dalam bentuk csv (*Comma Separated Values*).

4.2. *Text Preprocessing*

Tahapan *text pre-processing* yang dilakukan diantaranya sebagai berikut:

1. *Case folding*: penyeragaman bentuk huruf besar menjadi huruf kecil
2. *Remove* URL: *link* internet yang ada pada *tweet* akan dihapus.

3. *Unescape* HTML: menghapus file HTML dengan kata yang mengandung "<x> str </x>"
4. *Remove mention*: menghilangkan kata yang terdapat symbol "@".
5. *Remove number*: menghapus angka yang ada pada dokumen teks
6. *Remove punctuation*: semua tanda baca yang terdapat pada *tweet* akan dihapus
7. *Strip white space*: spasi yang berjumlah lebih dari satu pada *tweet* akan dihapuskan
8. Normalisasi Kata: mengubah kata tidak baku menjadi kata yang ada pada KBBI.

4.3. *Feature Selection*

Pada tahap *stopwords removal* digunakan kamus *stopwords* sebanyak 758 kata dan ditambah kamus *stopwords* manual sebanyak 2.187 kata. Tahap *stemming* menghilangkan imbuhan kata yang ada dengan menggunakan package katadasaR. Sedangkan pada tahap *tokenizing* membagi kalimat menjadi potongan kata yang tidak saling berpengaruh yang dipisahkan menggunakan *white space* atau *space*.

4.4. *Text Representation*

Setelah melalui tahap *text preprocessing* dan *feature selection*, jumlah data *tweets* yang awalnya berjumlah 1164 berubah menjadi 931 data *tweets*. Pada proses ini, data *tweets* diubah menjadi matriks yang berisi frekuensi kemunculan kata (TF) pada sebuah dokumen, serta dilakukan pembobotan menggunakan pembobotan TF-IDF. Berikut hasil nilai pembobotan kata menggunakan TF-IDF yang dapat dilihat pada Tabel 1.

Tabel 1. Pembobotan Kata dengan TF-IDF

No	*Tweet*	aksesori	cashback	dapat	diskon	⋯	gadget	promo
8	aksesori gadget gadget dapat diskon cashback	1,306	0,577	0,601	0,513	⋯	1,730	0
31	buku harga ribu promo serba ribu serbaserbi	0	0	0	0	⋯	0	0,434
119	dapat diskon cashback diskon elektronik belanja	0	0,577	0,601	1,027	⋯	0	0
258	promo gajianseru voucher	0	0	0	0	⋯	0	1,013
592	buku dapat buku favorit harga spesial	0	0	0,601	0	⋯	0	0

4.5. Pengaplikasian Algoritma *Fuzzy C-Means* dan *Fuzzy Possibilistics C-Means*

4.5.1. *Fuzzy C-Means*

Proses *clustering* menggunakan algoritma *Fuzzy C-Means* di uji coba dari *cluster* 3 sampai 7.

Tabel 2. Hasil Proses *Clustering* dengan algoritma FCM

Cluster	Waktu Komputasi	Iterasi	Fungsi Objektif	*Modified Partition Coefficient*
3	0,81	9	4567,600	4.884981×10^{-15}
4	1,0	8	3425,700	9.547918×10^{-15}
5	1,4	9	2740,560	1.332268×10^{-15}
6	1,67	9	2283,800	1.221245×10^{-15}
7	2,09	10	1957,543	1.110223×10^{-15}

4.5.2. *Fuzzy Possibilistics C-Means*

Proses *clustering* menggunakan algoritma *Fuzzy Possibilistics C-Means* di uji coba dari cluster 3 sampai 7.

Tabel 3. Hasil Proses *Clustering* dengan algoritma FPCM

Cluster	Waktu Komputasi	Iterasi	Fungsi Objektif	*Modified Partition Coefficient*
3	0,21	2	4567,624	-0.4999849
4	0,27	2	3425,731	-0.3333155
5	0,34	2	2740,599	-0.2499791
6	0,45	2	2283,846	-0.1999759
7	0,52	2	1957,597	-0.1666393

Berdasarkan Tabel 2 dan 3 dapat dilihat bahwa dari kedua metode yang memenuhi kriteria waktu komputasi dan jumlah iterasi adalah *Fuzzy Possibilistics C-Means*, sedangkan untuk metode yang memenuhi kriteria fungsi objektif minimum dan indeks *Modified Partition Coefficient* adalah *Fuzzy C-Means*. Sehingga rekomendasi metode *clustering* terbaik dalam kasus ini adalah *Fuzzy C-Means*.

Menurut hasil percobaan yang dilakukan, ada beberapa hal yang dapat dianalisis. Untuk menentukan *cluster* mana yang memiliki nilai validitas terbaik, digunakan indeks *Modified Partition Coefficient* (MPC) untuk setiap *cluster* yang diuji pada algoritma *Fuzzy C-Means*.

Tabel 4. Hasil Perhitungan Jumlah *Cluster* Optimum

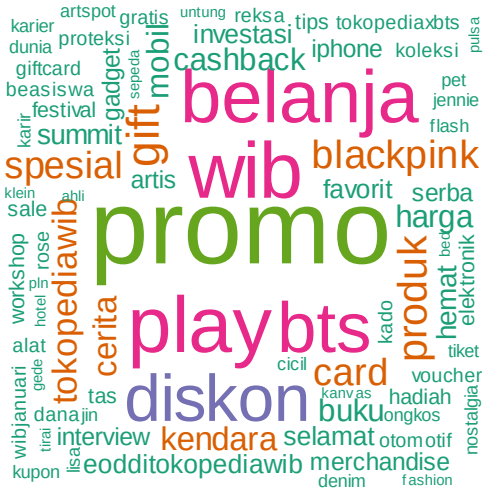
Cluster	*Modified Partition Coefficient*
3	4.884981×10^{-15}
4	9.547918×10^{-15}
5	1.332268×10^{-15}
6	1.221245×10^{-15}
7	1.110223×10^{-15}

Berdasarkan Tabel 4 dapat terlihat bahwa jumlah *cluster* yang optimal adalah 4 *cluster* karena memiliki nilai indeks *Modified Partition Coefficient* (MPC) terbesar sebesar 9.547918×10^{-15}. Oleh karena itu, jumlah *cluster* yang akan dianalisis dalam penelitian ini adalah 4 *cluster*. Hasil *clustering* dengan jumlah 4 *cluster* adalah sebagai berikut.

Tabel 5. Hasil *Clustering* dengan Algoritma Fuzzy C-Means

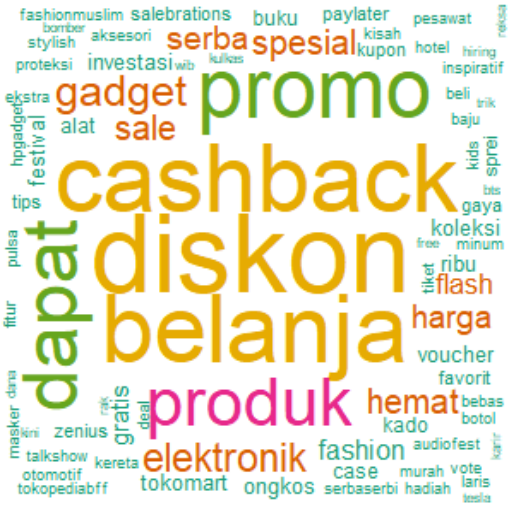
Cluster	Nomor Anggota *Tweet*	Jumlah Anggota
1	1, 2, 3, 4, 6, 7, 12, 13, 14, 16, 17, 18, 20, 23, 24, 25, 27, 28, 30, 33, 34, 44, 46, 47, 49, 50, 51,53, …, 898	274
2	5, 8, 10, 21, 31, 39, 40, 42, 43, 45, 48, 52, 54, 57, 60, 65, 66, 70, 71, 74, 74, 85, 87, 90, 96, 100, …, 910	236
3	9, 15, 22, 32, 35, 36, 55, 56, 63, 64, 73, 82, 91, 95, 102, 106, 112, 115, 122, 133, 134, 136, …, 913	185
4	11, 19, 29, 37, 38, 41, 58, 59, 61, 62, 72, 75, 76, 86, 89, 93, 97, 108, 110, 117, 143, 151, 163, …, 903	218

Dalam menentukan jenis konten dari setiap *cluster tweet*, harus dicari terlebih dahulu kata yang paling sering muncul pada setiap *cluster*. Proses pencarian kata ini dapat dilakukan dengan melihat *wordcloud* yang telah dibentuk dari masing-masing *cluster tweets*.



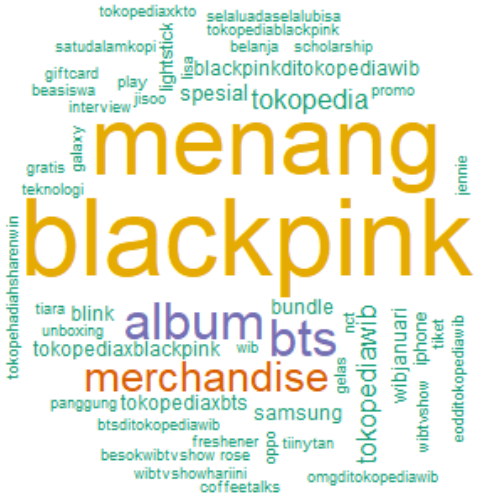
Gambar 1. *Wordcloud* untuk *Cluster* 1

Dari *wordcloud* pada Gambar 1 terlihat bahwa pada *cluster* 1 kata dengan frekuensi muncul paling tinggi adalah "promo", "wib", dan "play". Setelah ditinjau kembali dengan mencari *tweet* yang berisi kata-kata berwarna hijau dan ungu dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 1 adalah *tweet* yang berisi berbagai penawaran menarik barang-barang yang bisa didapatkan secara khusus pada *streaming platform* yang dimiliki oleh Tokopedia yang bernama Tokopedia Play. Seperti barang otomotif, alat olahraga, dan *merchandise* artis ternama.



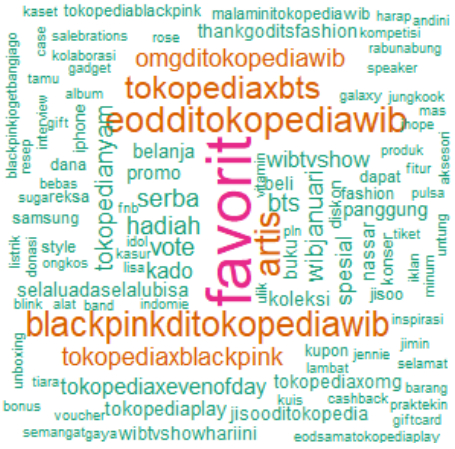
Gambar 2. *Wordcloud* untuk *Cluster* 2

Dari *wordcloud* pada Gambar 2 terlihat bahwa *cluster* 2 yang memiliki frekuensi kemunculan paling tinggi adalah "*cashback*", "diskon", dan "belanja". 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, *cashback*, dan belanja yang bisa dipakai saat berbelanja di Tokopedia. Seperti diskon tiket pesawat, alat elektronik, dan *gadget*. Kata diskon di dalam *cluster* 2 muncul dalam 81 *tweets*, sedangkan untuk kata *cashback* muncul dalam 73 *tweets* dari 236 *tweets*.



Gambar 3. *Wordcloud* untuk *Cluster* 3

Dari *wordcloud* pada Gambar 3 terlihat bahwa pada *cluster* 3 kata yang paling banyak muncul adalah "blackpink", "menang", "bts", dan "album". 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 ucapan selamat kepada para pemenang kuis berhadiah album Blackpink dan *merchandise* BTS. Kata "blackpink" di dalam *cluster* 3 muncul dalam 76 *tweets*, sedangkan kata "menang" di dalam *cluster* 3 muncul dalam 67 *tweets* dari 185 *tweets*.



Gambar 4. *Wordcloud* untuk *Cluster* 4

Dari *wordcloud* pada Gambar 4 terlihat bahwa pada *cluster* 4, kata yang memiliki frekuensi kemunculan paling tinggi adalah "favorit". Setelah ditinjau kembali dengan mencari *tweet* yang berisi kata berwarna ungu dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 4 adalah *tweet* yang berisi tentang penawaran barang yang paling sering dibeli ataupun dicari di Tokopedia. Kata "favorit" di dalam *cluster* 4 muncul dalam 33 *tweets* dari 218 *tweets*. Sedangkan untuk kata yang berwarna jingga adalah *hashtag* yang dipakai untuk mempromosikan acara yang diselenggarakan oleh Tokopedia di stasiun televisi atas terpilihnya *Brand Ambassador* Tokopedia yang terbaru.

4.6. Penentuan Konten yang Disukai Followers @tokopedia

Dengan menghitung rata-rata jumlah retweet dan *likes* untuk setiap *cluster*, dapat diketahui konten yang disukai oleh *followers* akun Twitter @tokopedia. Hal tersebut

dikarenakan jika pengguna Twitter menyukai *tweet* tersebut, pengguna dapat menekan tombol *retweet* atau *like*.

Tabel 6. Hasil Perhitungan Rata-rata Jumlah *Like* dan *Retweet*

Cluster	Jumlah *Tweet*	Jumlah *Like*	Jumlah *Retweet*	Rata-rata *Like*	Rata-rata *Retweet*
1	274	1176750	324138	4294,708	1182,985
2	236	229175	44451	971,081	188,352
3	185	742404	182331	4012,995	985,573
4	218	777170	159152	3565	730,055

Pada Tabel 6 terlihat bahwa *cluster* dengan rata-rata *like* dan *retweet* tertinggi adalah *cluster* 1, yaitu tentang *tweet* mengenai promo barang otomotif, alat olahraga, dan *merchandise* artis ternama. Lalu, jenis konten dengan rata-rata *like* dan *retweet* terendah adalah *cluster* 2, yaitu *tweet* mengenai penawaran tiket pesawat, alat elektronik, dan *gadget*. Berdasarkan hasil yang diperoleh, terlihat bahwa konsumen perusahaan *e-commerce* PT Tokopedia lebih tertarik penawaran mengenai barang-barang otomotif, alat olahraga, dan *merchandise* artis ternama yang disajikan dalam *streaming platform* dalam fitur Tokopedia Play dibandingkan dengan penawaran mengenai tiket pesawat, alat elektronik, dan *gadget*. Oleh karena itu, PT Tokopedia diharapkan dapat meningkatkan konten penawaran mengenai barang-barang otomotif, alat olahraga, dan *merchandise* artis ternama yang bisa didapatkan di Tokopedia ataupun melalui fitur Tokopedia Play sebagai sarana *advertising* semaksimal mungkin.

5. KESIMPULAN

Hasil perbandingan metode *Fuzzy C-Means* (FCM) dan *Fuzzy Possibilistics C-Means* (FPCM) pada studi kasus data *tweets* akun Twitter @tokopedia dengan menggunakan indeks validitas *cluster Modified Partition Coefficient* (MPC) ialah metode *Fuzzy C-Means* merupakan metode yang lebih baik dibandingkan dengan metode *Fuzzy Possibilistics C-Means*. Penerapan algoritma *Fuzzy C-Means* untuk *clustering tweets* pada akun Twitter @tokopedia menghasilkan 4 *cluster tweets* dengan konten mengenai penawaran barang promo dan diskon, *cashback*, kuis berhadiah serta mengenai promosi acara yang diselenggarakan oleh Tokopedia. Didapatkan jenis konten dengan *rata-rata* jumlah *like* dan *retweet* tertinggi yaitu mengenai promo barang otomotif, alat olahraga, dan penawaran *merchandise* artis ternama, dan untuk rata-rata terendah yaitu mengenai penawaran tiket pesawat, alat elektronik, dan *gadget*. PT Tokopedia dapat mengetahui konten *tweet* yang banyak disukai dan yang dapat menarik perhatian oleh para *followers* @tokopedia, sehingga informasi ini dapat digunakan sebagai sarana *advertising* pada *platform* media sosial Twitter.

DAFTAR PUSTAKA

Castella, Q. & Sutton, C., 2014. *Word Storms: Multiples of Word Clouds for Visual Comparison of Documents.* Seoul, s.n.

Chaudhuri, A., 2015. Intuitionistic Fuzzy Possibilistic C Means Clustering Algorithms. *Advances in Fuzzy Systems,* pp. 1-17.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook: Advances Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Hanifah, R. & Nurhasanah, I. S., 2018. Implementasi Web Crawling untuk Mengumpulkan Informasi Wisata Kuliner di Bandar Lampung. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),* Oktober, Volume 5, No. 5, pp. 531-536.

Kusumadewi, S. & Purnomo, H., 2013. *Aplikasi Logika Fuzzy: Untuk Pendukung Keputusan.* 2 ed. Yogyakarta: Graha Ilmu.

McNaught, C. & Lam, P., 2010. Using Wordle as a Supplementary Research Tool. *The Qualitative Report,* May, Volume 15 Number 3 , pp. 630-643.

Nurjannah, M., H. & Astuti, I. F., 2013. Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) untuk Text Mining. *Jurnal Informatika Mulawarman,* Volume 8, No. 3, pp. 110-113.

Siyamto, Y., 2017. Pemanfaatan Data Mining dengan Metode Clustering. *Media Informatika Budidarma,* Volume 1 No 2, pp. 28-31.

Suleman, A., 2015. A new perspective of modified partition coefficient. *Pattern Recognition Letters,* Volume 56, pp. 1-6.

Whiting, A. & Williams, D., 2013. Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal,* Volume 16 No. 4, pp. 362-369.

Yan, J., 2009. Text Representation. *Encyclopedia of Database System*.