ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 11, Nomor 1, Tahun 2022, Halaman 118 - 129
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENGELOMPOKAN *TWEETS* PADA AKUN TWITTER TOKOPEDIA MENGGUNAKAN ALGORITMA *DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE*

Deanira Qinanty Alamsyah[1], Sudarno[2*], Puspita Kartikasari[3]
[1,2,3]Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***email:** dsghani@gmail.com

ABSTRACT

Social media has become a trend for Indonesian people to express opinions, socialize, and exchange ideas. Internet users in Indonesia in 2021 will reach 202.6 million, 84% of whom use the internet to access social media. Twitter is one of the popular social media in Indonesia. This phenomenon is an opportunity for companies to use Twitter as a marketing tool, one of which is a marketplace company in Indonesia, Tokopedia. This research is intended to cluster tweets uploaded by the @tokopedia Twitter account to find out the type of content that gets a lot of likes and retweets by followers of the @tokopedia Twitter account. Cluster formation is done by applying the Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN). DBSCAN is a clustering algorithm based on density. The DBSCAN algorithm requires two parameters, namely the radius (Eps) and the minimum number of objects to form a cluster (MinObj). This research conducted several experiments with different Eps and MinObj parameters on 1.344 tweets that had gone through the stages of removing duplication, text preprocessing, and feature selection. The quality of the cluster formed is measured using the Silhouette Coefficient. Based on the highest average Silhouette Coefficient, the parameter values of Eps=5 and MinObj=3 with Silhouette Coefficient = 0.575 are determined as the best parameters that produce 2 clusters and 7 noise. The type of content that has the highest average number of likes and retweets is the WIB (Indonesian Shopping Time) campaign, so Tokopedia can use this type of content as a marketing tool on Twitter social media because this type of content is preferred by followers of the @tokopedia Twitter account.

Keywords: Twitter, Tokopedia, *Clustering*, DBSCAN, *Silhouette Coefficient*

1. PENDAHULUAN

Media sosial kini telah menjadi *trend* bagi masyarakat Indonesia untuk mengungkapkan pendapat, bersosialisasi, serta bertukar pikiran. Hingga Januari 2021 pengguna internet di Indonesia mencapai 202,6 juta orang dimana 84 persen diantaranya menggunakan internet untuk mengakses media sosial[2]. Dengan kemampuan penyebaran serta kemudahan akses tersebut, media sosial memberikan alternatif baru bagi perusahaan untuk melakukan pemasaran produk atau jasa yang biasa disebut dengan *social media marketing.* Twitter merupakan salah satu media sosial yang banyak digunakan untuk kegiatan *social media marketing*.

PT Tokopedia atau yang lebih dikenal dengan Tokopedia merupakan salah satu perusahaan jual beli *online* di Indonesia yang menggunakan Twitter sebagai media pemasaran. Tokopedia berhasil menduduki peringkat pertama *e-commerce* paling top di Indonesia dengan nilai *monthly traffic estimated* sebesar Rp 148.500.000,00. PT Tokopedia pada awal tahun 2021 berhasil mempertahankan eksistensi di media sosial Twitter dengan memperoleh jumlah pengikut sebanyak 808 ribu. Jumlah ini merupakan jumlah terbanyak dibandingkan dengan *e-commerce* lainnya[1]. Tokopedia dengan nama akun Twitter @tokopedia banyak memberikan informasi mengenai promo yang sedang ditawarkan,

produk-produk yang dijual, kuis berhadiah, serta penawaran-penawaran menarik lainnya. Informasi-informasi tersebut diberikan guna menarik audiens untuk melakukan transaksi di Tokopedia. Oleh karena itu, penting bagi Tokopedia untuk mengetahui jenis konten seperti apa yang disukai audiens agar dapat menentukan strategi pemasaran yang tepat.

Metode yang dapat digunakan untuk menganalisa data Twitter adalah *text mining*[3]. Salah satu tahapan lanjutan dari metode *text mining* yaitu *clustering*. Salah satu metode *clustering* yang dapat digunakan adalah DBSCAN (*Density Based Spatial Clustering of Applications with Noise*). DBSCAN adalah metode *clustering* yang mengelompokan data berdasarkan densitas data yang terkoneksi (*density connected*) [8]. Algoritma DBSCAN unggul dengan kemampuannya dala mendeteksi *outlier/noise.* Algoritma ini membutuhkan dua parameter yang harus ditentukan dengan tepat, yaitu Epsilon (ε) dan minimum objek (MinObj) [12]. Dalam penelitian ini akan menggunakan *Silhouette Coefficient* dalam mengukur validitas *cluster* yang terbentuk.

Penelitian ini akan mengelompokan *tweets* dari akun Twitter @tokopedia menggunakan algoritma DBSCAN (*Density Based Spatial Clustering of Applications with Noise*). Data yang digunakan adalah *tweet* yang diunggah akun Twitter @tokopedia serta jumlah *like* dan *retweet* dari masing-masing *tweets* yang diunggah untuk mengetahui *tweets* yang paling disukai oleh *followers* akun Twiter @tokopedia menggunakan *software* RStudio.

2. TINJAUAN PUSTAKA

2.1. *Social Media Marketing*

Social media markerting sebagai kegiatan mendorong individu untuk melakukan promosi terkait situs web, produk dan layanan mereka melalui saluran sosial *online* dan untuk berkomunikasi dengan cara memanfaatkan komunitas yang jauh lebih besar sehingga memiliki kemungkinan lebih besar untuk melalukan pemasaran daripada melalui saluran periklanan konvensional[13].

2.2. *Text Mining*

Text mining merupakan 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 [3]. Tahap-tahap *text mining* adalah sebagai berikut [3]:

1. *Text Pre-Processing*

   Text pre-processing meliputi berbagai jenis teknik ekstraksi informasi yang mengubah format mentah, tidak terstruktur, dan memiliki format asli menjadi terstruktur dan dapat diolah pada tahapan berikutnya [3]. Tahap-tahap *preprocessing* yang dilakukan antara lain:

   a. *Case Folding*, yaitu mengkonversi keseluruhan teks dalam dokumen menjadi huruf kecil.
   b. *Remove* URL, yaitu menghapus *link* internet (*Uniform Resources Locator*).
   c. *Unescape* HTML, yaitu menghilangkan file HTML serta jejak karakter yang dapat diduga sebagai *markup language*.
   d. *Remove Mention*, yaitu menghilangkan rujukan kepada pengguna Twitter lain yang diawali dengan simbol "@".

e. *Remove Number*, menghilangkan angka yang terdapat dalam teks.
f. *Remove Punctuation*, menghapus karakter *non alphabet* yang biasanya berupa tanda baca.
g. *Remove Emoticon*, menghilangkan simbol *emoticon* yang ada pada teks.
h. *Strip WhiteSpace*, menghapus spasi yang berlebih pada dokumen.
i. Normalisasi Kata, mengubah kata tidak baku menjadi kata baku sesuai dengan KBBI.

2. *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 [3]. Proses yang dilakukan pada tahapan ini adalah:

1. *Stopwords Removal*, pembuangan kata-kata yang sering muncul namun tidak relevan dalam suatu dokumen.
2. *Stemming*, pengubahana kata berimbuhan menjadi kata dasar.
3. *Tokenizing*, membagi rangkaian kata dalam dokumen berdasarkan spasi.

3. *Text Representation*

Text representation adalah tahapan mengubah data tekstual menjadi representasi yang lebih mudah untuk diproses. Salah satunya menggunakan *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*). Setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, sehingga perlu dilakukan pembobotan untuk setiap kata yang digunakan. TF-IDF adalah cara untuk memberikan bobot hubungan suatu kata atau *term* terhadap suatu dokumen [11]. TF-IDF diformulasikan oleh persamaan berikut [3]:

$$W_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} . log_2 \frac{D}{d_i} \quad (1)$$

dengan

$W_{i,j}$: Pembobotan TF-IDF untuk *term* ke-*i* pada dokumen ke-*j*
i : Banyaknya *term* (1,2,...,p)
j : Banyaknya dokumen (1,2,...,N)
$n_{i,j}$: Banyak kemunculan *term* ke-*i* pada dokumen ke-*j*
$\sum_k n_{k,j}$: Jumlah kemunculan seluruh *term* pada dokumen ke-*j*
D : Banyak keseluruhan dokumen
d_i : Banyaknya dokumen yang mengandung *term* ke-*i*.

2.3. *Clustering*

Clustering didefinisikan sebagai suatu proses pengelompokan data ke dalam sebuah kelompok (*cluster*), dimana objek yang berada di dalam suatu *cluster* memiliki tingkat kemiripan (homogenitas) yang tinggi satu sama lainnya tetapi memiliki tingkat ketidakmiripan (heterogenitas) yang tinggi dengan objek di *cluster* lain[5]. *Clustering* mengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Oleh karena itu,

clustering digolongkan sebagai metode *unsupervised learning*[6].

2.4. Jarak Euclidean

Memilih ukuran jarak merupakan hal yang harus dilakukan dalam analisis *cluster* karena karena tujuan dari *clustering* adalah mengelompokan objek-objek yang memiliki kemiripan ke dalam satu kelompok. Salah satu pengukuran jarak yang dapat digunakan adalah Jarak Euclidean yang diformulasikan oleh persamaan berikut[9]:

$$d(x,y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \qquad (2)$$

2.5. *Density Based Spatial Clustering of Applications with Noise* (DBSCAN)

DBSCAN adalah metode *clustering* yang bekerja dengan cara mengelompokan data berdasarkan densitas yang terkoneksi (*density connected*), dimana daerah dengan densitas tinggi dianggap sebagai *cluster* sedangkan untuk densitas rendah dianggap tidak tergabung dalam *cluster* atau dianggap sebagai *noise*[8]. Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data (minObj) yang berada dalam radius Eps (ε) dari setiap data. DBSCAN menghasilan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise*[9]. Data dikatakan sebagai inti jika jumlah data tetangga dan dirinya sendiri pada radius ε berjumlah ≥ MinObj. Data disebut sebagai batas (*border*) jika jumlah tetangga dan dirinya sendiri dalam radius ε kurang dari MinObj, tetapi tetangganya menjadi inti karena kehadirannya. Jika jumlah tetangga dan dirinya sendiri dalam radius ε kurang dari MinObj dan tidak ada tetangga yang menjadi inti karena kehadirannya maka data tersebut disebut sebagai *noise*.

Pengelompokan dengan menggunakan DBSCAN secara umum dilakukan dengan algoritma sebagai berikut[12]:

1. Tandai semua D sebagai "*unvisited*"
2. Pilih sebuah objek *p* kemudian tandai sebagai "*visited*"
3. Uji apakah dalam radius ε objek *p* memiliki minimal MinObj objek tetangga
4. Jika iya sebuah *cluster* C akan terbentuk, tambahkan objek *p* ke *cluster* C
5. Jadikan N sebagai objek-objek dalam ε-*neighborhood* dari *p*
6. Lakukan secara iteratif kepada objek yang masih bertanda "*unvisited*"
7. Jika sebuah objek dalam radius ε tidak memiliki minimal MinObj objek tetangga, maka ditandai sebagai *noise*.

2.6. Penentuan Eps (ε) dan MinObj

Eps dan Minobj yang optimal dapat ditentukan dengan melihat grafik *k-dist* dengan melihat pergeseran nilai Eps dari nilai *k* yang bervariasi. Titik dimana akan terjadi perubahan tajam pada k-dist akan digunakan sebagai Eps dan nilai k akan digunakan sebagai MnObj [10].

2.7. Pengukuran Kualitas *Cluster* dengan *Silhouette Coefficient*

Silhouette Coefficient merupakan sebuah metode yang digunakan untuk mengatahui seberapa baik *cluster* yang terbentuk. Pengujian dilakukan dengan menghitung nilai $a(i)$ dan $b(i)$. $a(i)$ merupakan rata-rata jarak objek *i* dengan semua objek lain dalam

cluster tersebut. Sedangkan $b(i)$ adalah rata-rata jarak minimum dari objek *i* dengan semua objek pada *cluster* lain (yang bukan *cluster i*) [7]. Setelah didapat nilai $a(i)$ dan $b(i)$ kemudian dicari *Silhouette Coeffiient* menggunakan persamaan (2).

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (3)$$

Nilai *Silhouette Coefficient* berkisar antara -1 hingga 1 ($-1 \leq SC \leq 1$). Nilai yang dihasilkan menunjukkan kekuatan sebuah objek berada di suatu *cluster*. Jika nilai *Silhouette Coefficient* mendekati 1 berarti *cluster* yang berisi objek *i* sangat padat dan objek *i* terpisah jauh dari *cluster-cluster* lain. Sebaliknya, jika *Silhouette Coefficient* mendekati -1, hal ini berarti *cluster* yang berisi objek *i* tidak padat dan objek *i* sangat dekat (bahkan tumpang tindih) dengan *cluster-cluster* lain[12].

3. METODOLOGI PENELITIAN

3.1. Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini berupa teks yaitu data *tweets* dari beranda akun Twitter @tokopedia yang bukan termasuk *tweet* balasan kepada pengguna lain. Pengambilan data dilakukan dengan cara *crawling* data di Twitter dengan bentuan Twitter API (*Application Programming Interface*) dengan data yang diambil berupa data *tweet,* jumlah *retweet* serta keterangan lain yang mendukung. Jumlah data maksimal yang dapat diambil dari akun Twitter @tokopedia menggunakan Twitter API sebanyak 1344 *tweets*.

3.2. Tahapan Pengolahan Data

Pengelompokan tweets ini dilakukan dengan bantuan perangkat lunak RStudio dan Microsoft Excel. Tahap-tahap yang dilakukan adalah sebagai berikut:

1. *Crawling* data *tweet* dari akun Twitter @tokopedia.
2. Melakukan *text pre-processing*. Proses yang dilakukan pada tahapan ini adalah *case folding, remove URL, unescape HTML, remove number, remove punctuation, remove emoticon, strip whitespace,* normalisasi kata.
3. Melakukan *feature selection*. Tahapan *feature selection* yang dilakukan yaitu proses *stopwords removal* dan *stemming,* dan *tokenizing.*
4. Pembobotan TF-IDF baris berupa nomor dari *tweet* dan kolom berupa seluruh kata penyusun data *tweet.* Matriks tersebut selanjutnya dikenal dengan istilah *document term matrix.*
5. Penerapan algoritma DBSCAN (*Density Based Spatial of Applications with Noise*):
   a. Menentukan parameter Eps (ε) dan MinObj.
   b. Mengambil titik *p* yang belum dikunjungi secara acak, kemudian menghitung jarak titik *p* dengan titik lainnya berdasarkan nilai Eps (ε) dan MinObj yang telah didapat menggunakan jarak Euclidean pada persamaan (2).
   c. Jika jumlah jarak yang dihitung lebih dari atau sama dengan Eps (ε), data ditandai sebagai inti (*core*) dan tetanggannya dibagai batas (*border*). Kemudian sebuah kelompok (*cluster*) baru akan mulai teerbentuk.
   d. Jika jumlah jarak kurang dari Eps (ε), data ditandai sebagai *noise.*
   e. Tandai data tersebut sebagai data yang telah dikunjungi.
   f. Mengulangi kembali langkah a sampai e sampai semua data telah diproses.
6. Mengukur kualitas *cluster* menggunakan *Silhouette Coefficient.*
7. Interpretasi hasil *cluster* yang telah terbentuk.

4. HASIL DAN PEMBAHASAN

4.1 *Crawling Twitter Data*

Penelitian ini menggunakan *software* RStudio untuk mengambil data dari Twitter. Sebelum melakukan *crawling* Twitter, harus dilakukan integrasi antara Twitter API dengan RStudio. Pada saat bergabung dengan Twitter API akan didapat empat kode berupa *consumer key*, *consumer secret*, *access token*, dan *access token secret*. Proses integrasi antara Twitter dengan RStudio dilakukan menggunakan fungsi: '`setup_Twitter_auth(api_key,api_secret,access_token, access_token secret)`'. Pengambilan data tweets menggunakan *package* TwitteR dengan fungsi '`tweets <- user_Timeline('tokopedia',n=3200,excludeReplies = TRUE)`'. Pada Tabel 1 adalah contoh data *tweets* pada akun Twitter @tokopedia.

Tabel 1. Contoh *Tweet* dari Akun Twitter @tokopedia

No.	*Tweet*	Tanggal *Tweet*	Jumlah *Like*	Jumlah *Retweet*
1	Buat rumah kamu makin nyaman dengan alat-alat rumah tangga dari Tokopedia Parents. Mulai dari Rp 20RB + Cashback hingga Rp 100RB, cek disini yukk > https://t.co/VnjgvTPEZG #TokopediaParents #MomNBaby	07/04/2021 08:48	126	4
2	Masih banyak banget makanan hits lainnya yang bisa kamu dapetin di #TokopediaNyam Pujasera, ada 10.000 menu ter-hits dan FlashSale 5x sehari mulai dari 10Ribu aja. Cek selengkapnya di https://t.co/QWXFiGWNsD	07/04/2021 08:10	37	0

4.2 Pengolahan *Tweets* dengan *Text Mining*

Sebelum dianalisis data *tweets* yang berbentuk *unstructure text* harus diolah terlebih dahulu melalui tahap *text pre-processing*, *feature selection* dan *text representation* untuk merubah bentuk teks menjadi suatu matriks angka yang digunakan saat proses analisis data.

a. *Text Pre-Processing*

Text pre-processing dilakukan untuk mempersiapkan data *tweets* yang akan digunakan untuk pengolahan di tahap selanjutnya. *Tweet* yang masih bersifat tidak terstruktur akan diolah menjadi suatu data tekstual yang memiliki format sama. Pada tahap ini URL, HTMl, *mention*, dan karakter lain selain alfabet pada seluruh akan dihapus. Selain itu, pada tahap ini juga dilakukan normalisasi kata yang tidak baku menjadi kata baku sesuai dengan KBBI.

Tabel 2. Contoh Hasil Proses *Text Pre-Processing*

Sampel	Hasil *Text Pre-Processing*
1	buat rumah kamu semakin nyaman dengan alat alat rumah tangga mulai dari rupiah ribu cashback hingga rupiah ribu cek disini yuk
2	selamat kepada memenangkan iphone merchandise selamat sekali lagi tunggu pesan dari admin ya

b. *Feature Selection*

Feature selection dilakukan dengan cara menghapus terms yang tidak dibutuhkan atau tidak relevan. Tahap pertama yang adalah *stopwords removal*, yaitu penghapusan kata yang tidak penting atau tidak relevan dalam dokumen. Selanjutnya *stemming* atau menghapus imbuhan yang terdapat dalam kata. Terakhir proses *tokenizing*, yaitu mengubah kalimat dokumen menjadi potongan kata.

Tabel 3. Contoh Hasil Proses *Feature Selection*

Sampel	Hasil *Feature Selection*
1	rumah nyaman alat alat rumah tangga rupiah ribu cashback rupiah ribu
2	selamat menang iphone merchandise selamat

c. *Text Representation*

Pada tahap ini data *tweets* yang telah dilakukan *text pre-processing* dan *feature selection* akan diubah menjadi *Document Term Matrix* dengan pembobotan TF dan pembobotan TF-IDF. Pembobotan TF berguna untuk melihat *terms* yang paling sering muncul, sedangkan pembobotan Tf-IDF digunakan untuk proses clustering dengan algoritma DBSCAN. Berdasarkan hasil *text representation*, kata yang menyusun 1344 *tweets* dari akun Twitter @tokopedia berjumlah 1246 kata. Seluruh kata tersebut akan menjadi variabel dari tiap *tweet*, dengan komponen dari matriks berupa jumlah dari *terms* pada tiap *tweet*. Proses perubahan tersebut dilakukan dengan menggunakan fungsi `'dtm.tf.idf <- weightTfIdf(m = dtm.tf, normalize = TRUE)'`.

Tabel 4. Hasil *Document Term Matrix* Pembobotan TF

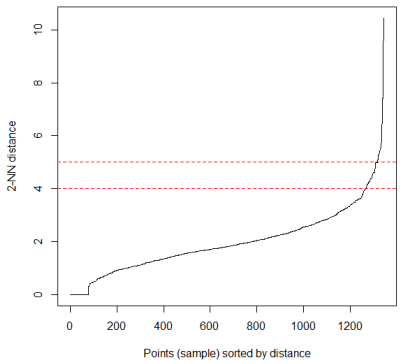
No.	*Tweet*	alat	cashack	diskon	...	menang	promo
82	diskon pesan tiket pesawat pesan kode promo cashback	0	1	1	...	0	1
243	selamat menang iphone merchandise selamat	0	0	0	...	1	0

Tabel 5. Hasil *Document Term Matrix* Pembobotan TF-IDF

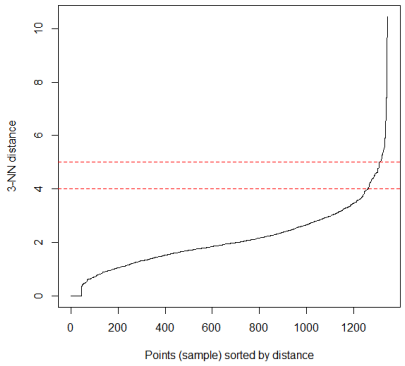
No.	*Tweet*	alat	cashack	diskon	...	menang	promo
82	diskon pesan tiket pesawat pesan kode promo cashback	0	0,291	0,257	...	0	0,321
243	selamat menang iphone merchandise selamat	0	0	0	...	0,655	0

4.1 Penentuan Eps dan MinObj

Pemilihan parameter Eps dan MinObj yang optimal ditentukan berdasarkan grafik *k-dist.* Komputasi dilakukan untuk mendapatkan nilai *k-dist* untuk selutuh titik pada $k = 2$ dan $k = 3$. Karena jika nilai k terlalu besar akan membentuk *cluster* yang kecil dan salah memberi label *noise*. Berdasarkan Gambar 1 dan Gambar 2, jika menggunakan MinObj = 2 maka nilai Eps yang optimal ada diantara 4 sampai dengan 5 dan jika menggunakan MinObj = 3 maka nilai Eps yang optimal berada diantara 4 sampai dengan 5.



Gambar 1. *K-dist* untuk $k = 2$



Gambar 2. *K-dist* untuk $k = 3$

4.2 Jarak Euclidean

Clustering tweets akun Twitter @tokopedia dibentuk menggunakan algoritma DBSCAN sehingga memerlukan jarak antar objek yang dihitung menggunakan jarak Euclidean. Hasil perhitungan jarak antar objek dapat dilihat pada Tabel 6.

Tabel 6. Perhitungan Jarak Euclidean

(i,j)	d(i,j)
(1,1)	0
(1,2)	3,737
⋮	⋮
(2,2)	0
(2,3)	5,784
⋮	⋮
(1343,1343)	0
(1343,1344)	3,537
(1344,1344)	0

4.3 *Clustering*

Hasil *clustering tweets* akun Twitter @tokopedia menggunakan algoritma DBSCAN dijabarkan pada Tabel 7.

Tabel 7. Hasil *Clustering Tweets* Akun Twitter @tokopedia

MinObj	Eps (ε)	Jumlah *Cluster*	*Noise*
2	4	4	60
	5	1	28
3	4	1	66
	5	2	7

4.4 Pengukuran Kualitas *Cluster*

Berdasarkan Tabel 8, dapat disimpulkan bahwa parameter terbaik untuk *clustering tweets* akun Twitter @tokopedia adalah MinObj = 3 dan Eps = 5 dan karena memiliki lebih sedikit *noise* dan memiliki *Silhouette Coefficient* yang paling besar dibandingkan parameter lainnya. Jadi, pada penelitian ini jumlah *cluster* yang digunakan ada 2 *cluster*.

Tabel 8. Pengukuran Kualitas *Cluster*

MinObj	Eps	Jumlah *Cluster*	*Noise*	*Silhouette Coefficient*
2	4	4	60	0,264
	5	1	28	0,465
3	4	1	66	0,382
	5	2	7	0,575

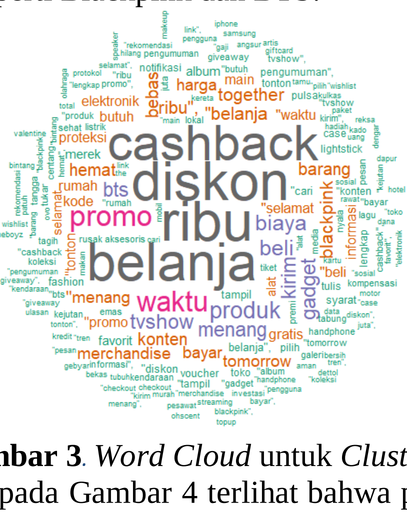
4.5 Interpretasi Hasil

Setelah diketahui jumlah cluster yang terbentuk, maka dilakukan interpretasi hasil *cluster*. Hasil *clustering* dengan jumlah 2 *cluster* dapat dilihat pada Tabel 9.

Tabel 9. Hasil *Clustering* dengan Algoritma DBSCAN

Cluster	Nomor Anggota *Tweet*	Jumlah Anggota
0	219, 390, 705, 721, 1014, 1114, 1161	7
1	1, 10, 12, 13, 14, 15, 16, 18, 22, 30, 33, 35, 37, 38, 40, 42, 43, 47, 49, 52, 53, 56, 57, 59, 60, 62, 64, 66, 69, 70, 73, 75, 76, 78, 79, 81, ... , 1343	744
2	2, 3, 4, 5, 6, 7, 8, 9, 11, 17, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 31, 32, 34, 36, 39, 41, 44, 45, 46, 48, 50, 51, 54, 55, 58, 61, 63, 65, 67, ... , 1344	593

Pada penelitian ini, ntuk menginterpretasikan lebih lanjut setiap *cluster* dibantu dengan *wordcloud*. Dari *wordcloud* Gambar 3 terlihat bahwa pada *cluster* 1 kata “diskon”, “belanja”, “*cashback*” dan “ribu” memiliki frekuensi kemunculan yang paling tinggi karena memiliki ukuran yang paling besar. Setelah ditinjau kembali dengan mencari *tweet* yang berisi kata-kata berwarna abu-abu dan ungu dapat disimpulkan bahwa *tweet* yang berada di dalam *cluster* 1 adalah *tweet* yang merujuk kepada kampanye WIB atau Waktu Indonesia Belanja yang diadakan Tokopedia setiap tanggal 25 hingga akhir bulan. Kampanye tersebut berisi ajakan untuk berbelanja di Tokopedia karena sedang diadakan diskon, gratis ongkos kirim (ongkir), serta *cashback* spesial. Selain itu, *tweets* pada *cluster* 1 membicarakan tentang WIB *TV Show* dimana acara tersebut menghadirkan penampilan spesial dari artis papan atas Internasional seperti Blackpink dan BTS.



Gambar 3. *Word Cloud* untuk *Cluster* 1

Dari *Word Cloud* pada Gambar 4 terlihat bahwa pada *cluster* 2 kata yang paling banyak muncul adalah “diskon”, “ribu”, “promo”, “produk”, dan “beli”. Kata berwarna hijau 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 diskon dan promo terhadap produk – produk yang dijual oleh Tokopedia seperti barang-barang elektronik, emas, makanan, buku, masker, dan lain sebagainya. Kata "diskon" dalam *cluster* 2 muncul sebanyak 83 kali dalam 593 *tweets*. Sedangkan kata "promo" muncul sebanyak 58 kali dalam 593 *tweets*.

Gambar 4. *Word Cloud* untuk *Cluster* 2

Dari *Word Cloud* pada Gambar 5 terlihat bahwa pada *noise* kata yang paling banyak muncul adalah "*giveaway*" dan "kejutan". Kata yang berwarna jingga merupakan kata dengan tingkat kemunculannya lebih rendah dibandingkan kata dengan warna lainnya.

Gambar 5. *Word Cloud* untuk *Noise*

Setelah ditinjau kembali dengan mencari *tweet* dengan kata-kata yang sering muncul dapat disimpulkan bahwa *tweet* yang berada di dalam *noise* adalah *tweet* yang berisi informasi mengenai *giveaway* dan kejutan - kejutan yang diadakan oleh Tokopedia. Kata "*giveaway*" di dalam *noise* muncul sebanyak 8 kali dalam 7 *tweets*, sedangkan untuk kata "kejutan" muncul sebanyak 5 kali dalam 7 *tweets*.

4.6 Penentuan Konten yang Disukai Followers Akun Twitter @tokopedia

Konten yang disukai oleh para *followers* akun Twitter @tokopedia 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 10. *Like* dan *Retweet* Masing-Masing *Cluster*

Cluster	Jumlah *Tweet*	Jumlah *Like*	Jumlah *Retweet*	Rata-rata *Like*	Rata-rata *Retweet*
0	7	11	1097	1587,56	156,7
1	744	1.678.360	436	2261,9	587,6
2	593	1.180.136	305	1990,1	514,4

Pada Tabel 29 terlihat bahwa *cluster* yang memiliki rata-rata *like* dan *retweet* tertinggi adalah *cluster* 1 yaitu *tweet* mengenai kampanye WIB atau Waktu Indonesia Belanja, sedangkan jenis konten yang memiliki rata-*rata retweet dan like* terendah adalah noise yaitu *tweet* mengenai *giveaway* yang diadakan oleh Tokopedia. Berdasarkan hasil yang didapatkan terlihat bahwa konsumen dari perusahaan *e-commerce* Tokopedia lebih tertarik dengan program WIB atau Waktu Indonesia Belanja yang diadakan setiap bulan oleh Tokopedia. Hal ini sesuai dengan riset perusahaan konsultan marketing MarkPlus Inc. pada tahun 2020 yang menyatakan bahwa kampanye WIB dari Tokopedia menempati peringkat kedua kampanye *e-commerce* yang paling dikenal masyarakat Indonesia sebanyak 71%. Oleh karena itu, PT Tokopedia Indonesia diharapkan mampu memanfaatkan konten WIB semaksimal mungkin.

5. KESIMPULAN

Penerapan algoritma DBSCAN untuk *clustering tweets* pada akun Twitter @tokopedia menghasilkan parameter yang optimal pada Eps = 5 dan MinObj = 3. Parameter tersebut menghasilkan 2 *cluster*, 7 noise, dan nilai *Silhouette Coefficient* sebesar 0,575. Dari dua *cluster* tersebut didapatkan jenis konten dengan rata-rata jumlah *retweet* dan *like* tertinggi yaitu *tweets* mengenai kampanye WIB (Waktu Indonesia Belanja), serta rata-rata terendah adalah *noise* mengenai informasi *giveaway*. Oleh karena itu, Tokopedia dapat menggunakan *tweets* dengan konten WIB sebagai sarana *marketing* pada media sosial Twitter karena *tweet* tersebut lebih disukai oleh para *followers* @tokopedia.

DAFTAR PUSTAKA

[1] Aseanup, 2019. *Top 10 E-commerce Sites in Indonesia 2019.* https://aseanup.com/top-e-commerce-sites-indonesia/. Diakses 28 Februari 2021.

[2] Datareportal, 2021. *Digital 2021: Indonesia.* https://datareportal.com/reports/digital-2021-indonesia. Diakses 2 Maret 2021.

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

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

[5] Han, J., Kamber, M., & Pei, J. 2012. *Data Mining Concept & Techniques*. Waltham: Elsevier Inc.

[6] Ian H. Witten & Eibe Frank. 2005. *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, San Francisco.

[7] L. Kaufman and P. J. Rousseuw. 1990. *Finding Groups in Data.* New York: John Wiley & Sons.

[8] Nagpal, P. B., & Mann, P. A. 2011. *Comparative study of density based clustering algorithms*. International Journal of Computer Applications, 27(11), 421-435.

[9] Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit ANDI.

[10] Purwanto, Barus, U. Y., Adrianto, B., & Agung, H. 2012. *Spatial Hotspots Clustering of Forest and Land Fires using DBSCAN and ST-DBSCAN*. Bogor.

[11] Robertson, S. 2005. *Understanding inverse document frequency: On theoretical arguments for IDF*. Journal of Documentation, Hal. 502-520.
[12] Suyanto, D. 2019. *Data Mining Untuk Klasifikasi dan Klasterisasi Data.* Bandung: Penerbit Informatika.
[13] Weinberg, T., 2009. *The New Community Rules : Marketing on the Social Web.* California: O'Reilly.