SMATIKA : STIKI Informatika Jurnal
Vol. 12, No. 1, Juni 2022, pp. 75~84
ISSN: 2087-0256, e-ISSN: 2580-6939

Clustering Data Tweet *E-Commerce* Menggunakan Metode K-Means (Studi Kasus Akun Twitter Blibli Indonesia)

Clustering E-Commerce Tweet Data Using the K-Means Method (Case Study of Blibli Indonesia's Twitter Account)

Alven Safik Ritonga[1*]
Isnaini Muhandhis[2]

[1,2]Teknik Informatika, Universitas Wijaya Putra, Surabaya
[1]alvensafik@uwp.ac.id, [2]isnainimuhandhis@uwp.ac.id

***Penulis Korespondensi:**
Alven Safik Ritonga
alvensafik@uwp.ac.id

Riwayat Artikel:
Diterima : 25 April 2022
Direview : 2 Juni 2022
Disetujui : 13 Juni 2022
Terbit : 27 Juni 2022

Abstrak

Perkembangan e-commerce sangat pesat pada saat ini, dengan semakin banyaknya e-commerce membuat persaingan dalam menarik customer dan mempertahankan loyal customer. Pelaku e-commerce perlu mencari strategi untuk hal tersebut, salah satu cara yaitu beriklan di sosial media, seperti;Twitter, Facebook, Instagram, dan lain sebagainya. Tujuan penelitian ini adalah mendapatkan clustering data tweet dari Twitter dengan menggunakan metode K-Means pada data tweet akun Twitter Blibli Indonesia untuk mengetahui jenis konten tweet yang banyak dilakukan retweet oleh followers. Data yang digunakan adalah data tweet follower yang ditarik dari akun Twitter @bliblidotcom. Pengujian jumlah cluster yang paling optimum dengan mencari nilai Silhouette coefficient yang terbesar. Hasil Penelitian diperoleh jumlah cluster yang optimal adalah 10 cluster. Dari hasil clustering ini diperoleh konten tweet yan paling disukai konsumen Blibli Indonesia adalah konten voucher(cluster 4) dan konten opporeno series (cluster 6). Konten voucher dan konten opporeno series hasil clustering ini bisa digunakan oleh Blibli untuk promo ke konsumennya.

Kata Kunci: Text Mining, Twitter, K-Means, E-commerce, Clustering

Abstract

The development of e-commerce is very rapid at this time, with the increasing number of e-commerce making competition in attracting customers and maintaining loyal customers. E-commerce players need to find a strategy for this, one way is advertising on social media, such as; Twitter, Facebook, Instagram, and so on. The purpose of this study was to obtain clustering of tweet data from Twitter using the K-Means method on tweet data from the Blibli Indonesia Twitter account to determine the type of tweet content that was retweeted by followers. The data used is follower tweet data which is pulled from the Twitter account @bliblidotcom. Testing the most optimum number of clusters by finding the largest Silhouette coefficient value. The results obtained that the optimal number of clusters is 10 clusters. From the results of this clustering, the tweet content that Blibli Indonesia consumers like the most is voucher content (cluster 4) and Opportunity series content (cluster 6). Voucher content and opporeno series content as a result of this clustering can be used by Blibli for promos to its consumers.

Keywords: Text Mining, Twitter, K-Means, E-commerce, Clustering

1. Pendahuluan

Salah satu media sosial yang banyak digunakan orang, bukan hanya untuk bicara politik, sosial, kemanusiaan, maupun hal yang umum, tapi juga membicarakan produk atau mencari barang yang diinginkan yaitu Twitter. Banyak orang berkunjung ke Twitter mencari produk dan menjadi pengikut (follower) pelaku usaha atau e-commerce, melakukan review produk maupun tanya jawab dengan e-commerce. Twitter menjadi tempat orang untuk membicarakan berbagai macam topik, termasuk belanja. Data Twitter memperlihatkan percakapan mengenai belanja meningkat 60 persen sejak Maret 2020 dibanding periode yang sama tahun sebelumnya. Menurut data Brandwatch, 44 persen pengguna Twitter di Indonesia berbicara mengenai belanja pakaian atau aksesori, makanan 40 persen, peralatan rumah, dan elektronik 35 persen, perawatan diri 33 persen, dan tentang ponsel atau gawai 27 persen. Sebanyak 41 persen masyarakat Indonesia di Twitter menemukan produk baru berdasarkan rekomendasi di media sosial. Seiring dengan semakin meningkatnya percakapan mengenai belanja di Twitter, pelaku usaha dapat memanfaatkan momentum ini untuk mempromosikan produk dan layanannya agar lebih banyak diketahui oleh konsumen.

Salah satu pelaku usaha atau e-commerce yang menggunakan Twitter sebagai media promosi adalah Blibli.com dengan nama akun Twitter @bliblidotcom. Blibli.com pada tahun 2020 peringkat 3 besar follower terbanyak di Twitter yaitu 514.800 follower. Blibli.com menggunakan Twitter untuk memberikan informasi produk-produk yang dijual, diskon, promo, kuis, dan beberapa tawaran untuk menarik minat pembeli supaya berbelanja di Blibli.com. Pelaku usaha terutama e-commerce harus bisa memilih konten yang di tweet untuk menarik minat followers, minimal menarik minat followers untuk retweet agar menyebarkan tweet yang dibuat oleh pelaku usaha [1].

Kebanyakan tweet yang dibagikan follower di Twitter berbentuk teks, maka untuk mendapatkan konten apa saja yang paling sering muncul dari tweet yang ada di Twitter digunakan metode text mining. Komentar (tweet) dan retweet yang dilakukan oleh followers berbentuk teks tersebut dapat dikelompokkan menjadi beberapa kluster berdasarkan kesamaan dan kemiripan konten [2]. Dalam penelitian ini metode untuk melakukan clustering adalah metode K-Means. Penentuan jumlah klaster terbaik dilakukan dengan menggunakan metode Silhouette coefficient. Peneliti menerapkan text mining dan Metode clustering K-Means untuk mengetahui kebiasaan orang memberikan komentar mengenai suatu konten e-commerce di sosial media Twitter. Hasil clustering yang diperoleh adalah kata apa yang paling sering digunakan atau diretweet oleh followers. Hasil ini bisa digunakan e-commerce untuk beriklan dan promo kepada konsumennya.

2. Metode Penelitian

Pada penelitian ini ada beberapa tahapan yang dilakukan sebagai berikut:

Pengumpulan Data

Pada tahapan pengumpulan data, data diambil dari Twitter yaitu data tweet, jumlah retweet, jumlah like, dan tanggal tweet yang ada di akun Twitter Blibli.com Indonesia yaitu @bliblidotcom. Cara mengambil data tweet dari Twitter menggunakan API (Application Programming Interface) dan R Studio. Yang kedua, pengambilan data dari data tweet yang diambil dari akun Twitter @bliblidotcom, mulai dari tanggal 18 Agustus 2021-26 Agustus 2021, jumlah data tweet yang diambil adalah 2164 tweet.

Teknik Pengolahan Data

Preprocessing Data

Case folding

Tahapan mengganti semua karakter huruf pada sebuah kalimat menjadi huruf kecil dan menghapus karakter angka, tanda baca, dan Uniform Resources Locator (URL).

Tokenizing
Pada tahap ini dilakukan pembagian tweet berdasarkan karakter spasi pada setiap kalimat.

Filtering
Tahapan menghapus tweet yang tidak dipakai dari proses sebelumnya. Cara untuk menghilangkan tweet dengan menggunakan kamus kata (stopword).

Stemming
Tahapan untuk mengganti semua tweet menjadi bentuk kata dasarnya.

Tagging
Tahapan ini biasanya dilakukan untuk teks yang bahasa Inggris atau bahasa lainnya. Pada tahap ini tujuannya adalah mengganti semua tweet dalam bentuk lampau menjadi kata awal.

Pembobotan Kata
Metode yang digunakan pada penelitian ini adalah Metode Term Frequency – Inverse Document Frequency (TF-IDF) yang merupakan cara pembobotan berdasarkan statistik yang sering diaplikasikan pada berbagai permasalahan penggalian informasi. Secara umum TF-IDF tidak banyak dikenal sebagai algoritma untuk peringkasan teks otomatis [3]. Pada peringkasan teks otomatis menggunakan TF-IDF, ide dasarnya adalah memberikan bobot pada setiap kalimat dalam sebuah dokumen. Setelah masing-masing kalimat diberikan bobot, kalimat akan diurutkan berdasarkan bobot dimana kalimat k teratas dengan bobot paling besar akan dipilih sebagai hasil akhir ringkasan. Bobot kalimat diperoleh dari penjumlahan bobot term pada sebuah kalimat, dimana term dapat berupa kata, frasa atau tipe sintatik lainnya. Berikut ini tahapannya [4], yaitu :

Term Frequency (tf)
Term frequency (tf) adalah jumlah seberapa sering suatu kata muncul pada suatu dokumen. Nilai W_{tf} adalah jumlah bobot nilai tf yang dihitung dengan menggunakan logaritma. Persamaan dari Term Frequency:

$$W_{tf_{t,d}} = \begin{cases} 1 + \log_{10} tf_{t,d} > 0 \\ 0, \quad otherwise \end{cases} \qquad (1)$$

Document Frequency (df)
Document Frequency (df) merupakan frekuensi atau jumlah dokumen yang mengandung suatu kata.

Inverse Document Frequency (idf)
Inverse Document Frequency (idf) adalah bobot nilai kebalikan dari nilai df. Kata yang frekuensi kemunculannya paling sedikit pada semua dokumen mempunyai bobot nilai Inverse Document Frequency yang tinggi. Persamaan dari Inverse Document Frequency (idf):

$$idf_t = \log_{10}\left({}^{N}/_{df_t}\right) \qquad (2)$$

Keterangan:
N : jumlah dokumen teks.
df_t : jumlah dokumen yang mengandung suatu kata t.

Term Frequency-Inverse Document Frequency (tf-idf)
Nilai bobot ini adalah perkalian nilai tf dan idf dari suatu kata, bentuk persamaannya berikut ini:

$$W_{t,d} = W_{tf_{t,d}} \times idf_t \qquad (3)$$

Keterangan:
$W_{tf_{t,d}}$:Term Frequency.

idf_t :Inverse Document Frequency.

Proses Clustering

Metode yang digunakan untuk clustering adalah metode K-Means. K-Means adalah metode yang tidak terlalu rumit dan tidak membutuhkan waktu yang lama dalam proses clustering obyek serta bisa mengelompokkan data dengan jumlah sangat besar. Langkah-langkah algoritma dari metode K-Means, berikut ini [5]. yang pertama, tentukan nilai k cluster. Yang kedua, bagi semua data ke dalam k cluster. Yang ketiga, hitung pusat cluster (sentroid) dari semua data di setiap cluster, rumusnya persamaan(4) berikut ini.

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \qquad (4)$$

nilai C adalah sentroid, M adalah jumlah data, i adalah jumlah cluster.

Keempat, semua data ditempatkan ke sentroid terdekat. Untuk menghitung jarak masing-masing data ke setiap sentroid digunakan rumus jarak Euclidean, rumusnya pada persamaan (5) berikut ini.

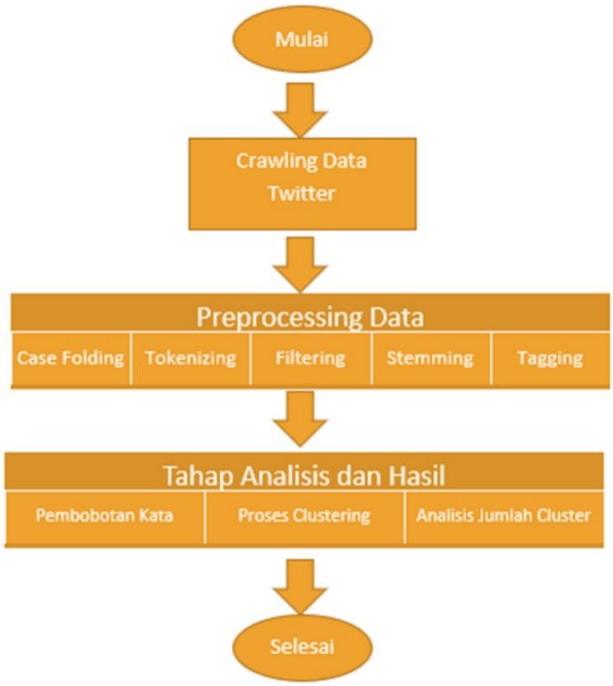
$$D(x_i, C_i) = \sqrt{\sum_{j=1}^{q} \left(x_{ij} - C_{ij}\right)^2} \qquad (5)$$

Kelima, Ulangi lagi langkah 3 jika masih ditemukan data yang berpindah cluster.

Analisis Jumlah Cluster

Menentukan jumlah cluster yang paling optimum dengan menghitung nilai Silhouette Coefficient, dengan mengkombinasikan nilai k (jumlah cluster) yang berbeda. Apabila nilai *Silhouette Coefficient* tertinggi itulah yang menjadi jumlah *cluster* paling optimum.

Tahapan penelitian bisa dilihat pada gambar 1 berikut ini.



Gambar 1. *Diagram Metode Penelitian*

3. Hasil dan Pembahasan

Data yang diambil dari akun Twitter @bliblidotdom dalam bentuk teks. Proses untuk menarik data dengan cara *crawling* dibantu oleh Twitter API, data yang didapat 2164 tweet mulai dari tanggal 18 Agustus 2021 sampai dengan tanggal 26 Agustus 2021. Data yang sudah diambil bisa dilihat pada Tabel 1 berikut ini.

Tabel 1. *Contoh Text Tweet*

No Tweet	Tanggal	Text Tweet	Jumlah Retweet
11	2021-08-26 20:53:13	RT @bliblidotcom: Sebut 1 kata untuk kamera selain..✨????????✨?? https://t.co/oTKLwf7SEG #BigPayDay https://t.co/UG8udHyvaW	2,0
12	2021-08-26 20:51:27	RT @bliblidotcom: CIVIC biru tuh.. ???? https://t.co/ApCMynFx17 #BlibliOTO https://t.co/EoIWygCsvP	5,0
13	2021-08-26 20:50:12	@bliblidotcom Promonya?? #NKKCHI https://t.co/Iu24clqj6y	,0
14	2021-08-26 20:49:29	RT @bliblidotcom: Kamu udah makan siang? Kl gt, #NKKCHI yuk! Nanti Kita Kasih Cepek Hari Ini! - FOLLOW aku, RT & LIKE tweet ini - Klik: ht…	59,0
15	2021-08-26 20:40:25	@bliblidotcom ????????. Lebih tasty sikit sih https://t.co/2uavM5JRHG	,0
16	2021-08-26 20:35:00	Tuker poin Blibli Rewards kamu dengan makanan yummmy?? https://t.co/Ps3tXe6me4 #BlibliINSPO https://t.co/ZeMRzkusfm	1,0

Sumber: Twitter Akun @bliblbidotcom

Pengolahan Data

Pada tahapan ini dimulai dengan teks preprocessing , prosesnya sebagai berikut:

Case Folding

Tahapan mengganti semua karakter huruf pada sebuah kalimat menjadi huruf kecil dan menghapus karakter angka, tanda baca, dan *Uniform Resources Locator* (URL). Hasil tahap ini bisa dilihat contoh di Tabel 2 berikut ini.

Tabel 2 *Contoh Teks Tweet Proses Case Folding*

No Tweet	Teks tweet yang Sudah Case Folding
11	kamera bigpayday
12	civic biru bliblioto
13	promonya
14	makan
15	tasty sikit
16	tuker poin blibli rewards makanan yummmy blibliinspo

Tokenizing

Tahap dimana sebuah kalimat dibagi menjadi beberapa kata yang membentuk kalimat itu. Contoh hasil tahap ini bisa dilihat pada tabel 3 berikut ini.

Tabel 3*.Contoh Hasil Proses Tokenizing*

No Tweet	Hasil Tokenizing
16	tukar poin blibli reward makanan yummmy blibliinspo

Filtering

Pada proses filtering ini dilakukan penghapusan kata yang tidak ada artinya dengan menggunakan kamus kata (stopword). Contoh proses filtering data tweet bisa dilihat pada tabel 4 berikut ini.

Tabel 4*. Contoh Hasil Proses Filtering*

No Tweet	Hasil Filtering
16	tukar poin blibli reward makanan

Pembobotan Kata

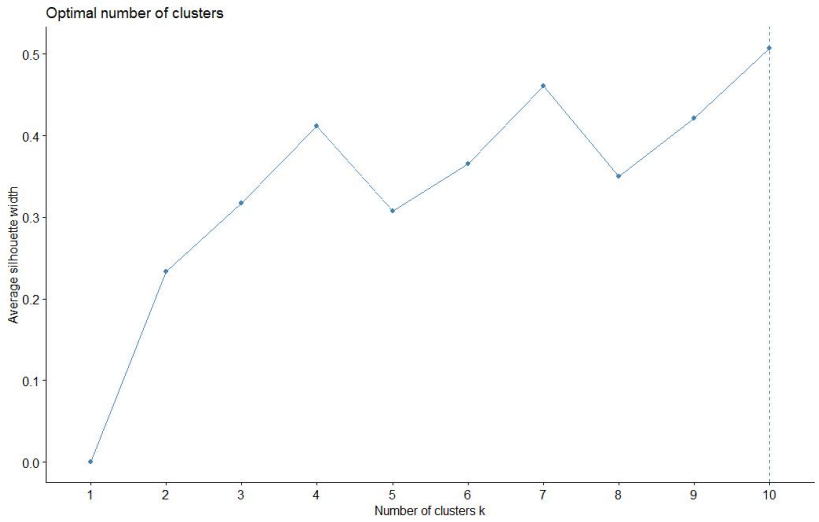
Prosesnya pembobotan menggunakan metode Term Frequency – Inverse Document Frequency (TF-IDF). Dengan cara merubah data tweet menjadi matriks yang setiap elemen matriks adalah jumlah kemunculan kata pada setiap data tweet. Dari hasil pembobotan diperoleh jumlah kata yang menyusun 2164 tweet adalah 155 kata. Contoh matriks proses pembobotan kata bisa dilihat pada tabel 5 berikut ini. Tabel yang didapatkan pada pembobotan ini yang akan dipakai untuk proses clustering dengan metode K-Means.

Tabel 5*.Contoh Hasil Pembobotan*

No Tweet	launching	liga	lokal	makan	menang	menyambut	nilai	nonton	opporeno	pantau
567	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.447	0.0
568	0.0	0.447	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
569	0.378	0.0	0.0	0.0	0.0	0.378	0.0	0.0	0.378	0.0
570	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
571	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
572	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Menentukan Jumlah Cluster Optimum

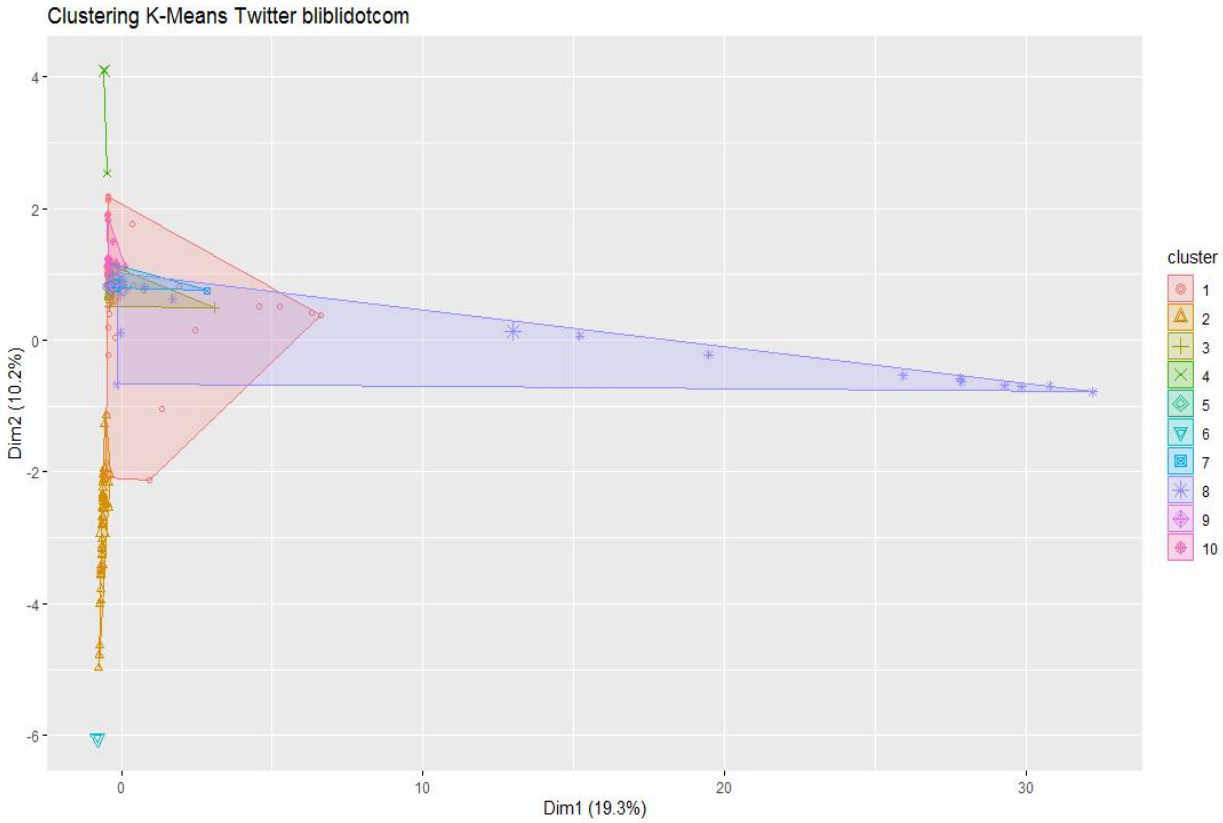
Pada metode K-means ini penentuan jumlah cluster optimum atau terbaik diambil dari cluster yang memiliki nilai Silhouette Coefficient tertinggi yaitu 0,5 dengan jumlah cluster k=10, bisa dilihat pada gambar 2 berikut ini.



***Gambar 2**. Grafik Silhouette Coefficient dengan Jumlah Cluster 1 sampai 10*

Proses Clustering dengan Metode K-Means

Mengunakan jumlah cluster optimum yaitu k=10, gambar cluster tweet akun bliblidotcom dalam bentuk grafik, dapat dilihat pada gambar 3 dibawah ini.



***Gambar 3**. Grafik Penyebaran Tweet bliblidotcom dalam 10 Cluster*

Hasil *clustering* di atas bisa ditentukan jumlah tweet setiap cluster dan jumlah retweet setiap cluster diperlihatkan pada tabel 6 di bawah ini.

***Tabel 6**. Hasil Clustering Menggunakan Metode K-Means*

Nomor Kluster	Jumlah Tweet	Jumlah Retweet	Rata-Rata Retweet
1	977	1314	1,34
2	121	6	0,05
3	213	296	1,39
4	156	5708	36,59
5	39	22	0,56
6	238	537	2,26
7	27	48	1,78
8	64	32	0,50
9	232	260	1,12
10	97	4	0,04

Berdasarkan tabel 6 di atas, bisa dianalisis bahwa ada beberapa cluster yang anggotanya tidak merata, misalnya cluster 1 dengan jumlah tweet yang lebih banyak dibandingkan cluster yang lain, ini disebabkan oleh banyak tweet yang sama pada cluster tersebut. Dari tabel 6 tersebut diperoleh bahwa ada beberapa cluster mempunyai rata-rata retweet yang tinggi yaitu cluster 4 dan cluster 6. Cluster yangg mempunyai rata-rata retweet yang rendah antara lain cluster 2, cluster 5, cluster 8 dan cluster 10. Konten yang sering muncul di retweet tertinggi yaitu voucher (cluster 4), konten promo opporene series (cluster 6).

Analisis Hasil Clustering dengan Metode K-Means

Pengujian kualitas cluster pada penelitian ini adalah berdasarkan nilai perhitungan Silhouette Coefficient.

***Tabel 7.** Nilai Sillhouette Coefficient Tiap Cluster*

Nomor Cluster	Jumlah Tweet	Nilai Sillhouette Coefficient
1	977	0,65
2	121	0,39
3	213	-0,12
4	156	0,99
5	39	0,35
6	238	1,00
7	27	0,59
8	64	0,14
9	232	0,75
10	97	0,17

Dari perhitungan nilai Sillhouette Coefficient Tiap Cluster diatas, dapat dilihat bahwa 9 cluster mempunyai nilai positif dan 1 cluster mempunyai nilai negatif. Hal ini menunjukkan bahwa pada 9 cluster yang bernilai positif, sebagian besar tweet pada cluster ini berada pada cluster yang sebenarnya. Dan sebaliknya 1 cluster yang nilai Silhouette Coefficient negatif menunjukkan sebagian besar tweet pada cluster tersebut tidak berada di cluster yang sebenarnya.

Analisis konten apa saja yang digunakan sebagai konten untuk beriklan diperoleh dari rata-rata jumlah retweet tiap cluster pada tabel 6 di atas. Didapatkan bahwa dari nilai retweet tertinggi yaitu konten voucher (cluster 4) dan konten promo opporeno series (cluster 6). Konten-konten ini divisualisasi dengan menggunakan wordcloud, seperti gambar visualisasi wordcloud di bawah ini.

***Gambar 4** .Visualisasi Wordcloud Cluster 4*

***Gambar 5**. Visualisasi Wordcloud Cluster 5*

Melihat analisis dan visualisasi di atas merperlihatkan bahwa konsumen Blibli lebih tertarik pada konten pemberian voucher dan promo opporeno series.

4. Penutup

Penerapan metode K-Means untuk mendapatkan hasil cluster dari data tweet e-commerce Blibli diperoleh 10 cluster. Untuk mendapatkan konten apa saja yang paling sering di retweet dan rata-rata retweet yang terbesar di setiap cluster, yaitu cluster 4 dan cluster 6. Dari kedua cluster tersebut diperoleh konten yang paling banyak diretweet di 2 cluster tersebut adalah voucher dan preorder opporeno series.

Blibli Indonesia bisa merekomendasikan hasil proses Clustering ini untuk mendata konten apa yang sering di ikuti oleh followernya di twitter, dan salah satu cara yang murah untuk beriklan dan promosi kepada konsumennya. Misalnya hasil cluster konten voucher dan preorder opporeno series, karena sering di retweet oleh follower bisa membantu Blibli untuk menambah follower atau konsumennya.

Pengembangan penelitian lanjutan dengan menggunakan beberapa metode clustering, supaya bisa dibandingkan metode apa yang cocok untuk clustering data tweet e-commerce Blibli.

5. Referensi

[1] D. S. Indraloka and B. Santoso, ”Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia,” Jurnal Sains dan Seni ITS, vol. 6, no. 2, pp. 51-56, 2017.

[2] R. K. Putri, B. Warsito and Mustafid, "Implementasi Algoritma Modified Gustafson-Kessel untuk Clustering Tweets pada Akun Twitter Lazada Indonesia, " Jurnal Gaussian, vol.8, no.3, pp. 285-295, 2019.
[3] D. H. Wahid and Azhari, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," IJCCS, vol. 10, no. 2, pp. 207-218, 2016.
[4] U. Rofiqoh, R.S. Perdana and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia pada Twitter dengan Metode Support Vector Machine dan Lexicon Based Features," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 1, no. 12, pp. 1725-1732, 2017.
[5] S. A. D. Budiman, D. Safitri and D. Ispriyanti,"Perbandingan Metode K-Means dan Metode DBSCAN Pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang," Jurnal Gaussian, vol. 5, no. 4, pp. 757-762, 2016.