ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 4, Tahun 2021, Halaman 583 - 593
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN *TEXT MINING* UNTUK MELAKUKAN *CLUSTERING* DATA *TWEET* AKUN BLIBLI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN *K-MEANS CLUSTERING*

Syiva Multi Fani[1], Rukun Santoso[2], Suparti[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
fani.syiva32@gmail.com

ABSTRACT

Social media is computer-based technology that facilitates the sharing of ideas, thoughts, and information through the building of virtual networks and communities. Twitter is one of the most popular social media in Indonesia which has 78 million users. Businesses rely heavily on Twitter for advertising. Businesses can use these types of tweet content as a means of advertising to Twitter users by Knowing the types of tweet content that are mostly retweeted by their followers . In this study, the application of Text Mining to perform clustering using the K-means clustering method with the best number of clusters obtained from the Silhouette Coefficient method on the @bliblidotcom Twitter tweet data to determine the types of tweet content that are mostly retweeted by @bliblidotcom followers. Tweets with the most retweets and favorites are discount offers and flash sales, so Blibli Indonesia could use this kind of tweet to conduct advertising on social media Twitter because the prize quiz tweets are liked by the @bliblidotcom Twitter account followers.

Keywords: Advertising, Blibli Indonesia, Clustering, K-means, Silhouette Coefficient, Text Mining, Twitter.

1. PENDAHULUAN

Twitter merupakan salah satu media sosial paling populer di Indonesia. Berdasarkan laporan *Wearesocial Hootsuite* pengguna media sosial Twitter di Indonesia mencapai 52% dari jumlah pengguna internet atau mencapai 78 juta pengguna [8]. Pengguna Twitter dapat mengirimkan pesan pendek dengan 140 karakter yang disebut dengan *tweet.* Twitter memiliki akses yang dinamakan *platform* API (*Application Programming Interface)* untuk memperoleh data dengan cepat dan dalam jumlah yang banyak untuk keperluan analisis [15].

Pelaku bisnis sangat mengandalkan Twitter untuk membantu memasarkan produk atau layanan yang mereka miliki agar dapat dikenal oleh konsumen. Blibli merupakan salah satu pelaku bisnis bidang *e-commerce* di Indonesia yang menggunakan media sosial Twitter sebagai sarana untuk melakukan periklanan. *Username* @bliblidotcom di Twitter telah memiliki jumlah *followers* sebanyak 501,3 ribu orang dan jumlah *tweet* sebanyak 65,3 ribu *tweets.* Dengan menemukan jenis konten *tweet* yang banyak dilakukan *retweet* oleh *followers* dari Blibli, diharapkan semakin banyak pengguna Twitter yang menjadi konsumen dari Blibli.

Pada penelitian ini dilakukan *clustering* menggunakan *software* R. Metode *clustering* yang digunakan adalah metode *K-means clustering* untuk mengelompokkan data tekstual berdasarkan kesamaan konten yang dimiliki ke dalam beberapa *cluster*. *Clustering* dilakukan pada data jenis konten *tweet* akun Twitter @bliblidotcom yang disukai oleh *followers* akun Twitter @bliblidotcom berdasarkan perhitungan rata-rata jumlah *favorite* dan *retweet* masing- masing *cluster*. Hal tersebut dikarenakan jika pengguna Twitter menyukai suatu *tweet* maka kemungkinan yang akan dilakukan pengguna tersebut adalah menekan tombol *favorite* atau melakukan *retweet*. Hasil *clustering* data *tweet* dapat digunakan untuk menunjang kegiatan *advertising* atau periklanan Blibli pada media sosial Twitter. Pelaku bisnis dapat memanfaatkan *followers* akun Twitter mereka sebagai sarana untuk melakukan *advertising*. Dengan menggunakan perintah *retweet* dan *favorite* pada Twitter, *followers* dapat menyebarkan *tweet* yang dibuat oleh pelaku bisnis. Dengan menemukan jenis konten *tweet* yang banyak disukai oleh *followers* dari @bliblidotcom, diharapkan semakin banyak pengguna Twitter yang menjadi konsumen dari Blibli Indonesia.

2. TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah sumber data melalui identifikasi pola yang menarik. Pada kasus *text mining*, sumber data adalah kumpulan data tekstual yang tidak terstruktur pada dokumen [3]. *Text mining* bertujuan untuk menemukan informasi yang tidak diketahui, sesuatu yang belum diketahui dan belum dapat ditulis [5].

2.2 Ekstraksi *Tweets*

Application Programming Interface (API) berfungsi sebagai penghubung antara sistem yang dibangun dengan twitter yang memungkinkan pengguna untuk mengekstrak data dengan pemrograman. Untuk dapat mengeksraksi *tweet* dari Twitter API diperlukan proses autentifikasi. Autentifikasi adalah suatu proses validasi atau pembuktian terhadap identitas seorang pengguna pada saat akan mengakses sebuah system [9]. Proses autentifikasi pada Twitter API ini menggunakan API *key*, API *secret key*, *access token*, dan *access token secret* yang dijadikan sebagai kode akses untuk memasuki sistem tersebut.

2.3 *Text Preprocessing*

Data *tweet* yang telah diambil dari sosial media twitter masih merupakan data mentah dengan karakteristik teks memiliki dimensi yang tinggi, terdapat *noise* pada data, dan terdapat struktur teks yang tidak baik[4]. Dalam *text preprocessing* dilakukan pengurangan kata-kata tidak penting, tidak mempunyai arti dari *database* teks atau dokumen, sehingga membuat data lebih terstruktur dan siap untuk diolah [7].

Berdasarkan penelitian yang dilakukan [17] langkah-langkah *text preprocessing* adalah sebagai berikut:

1. *Case folding* adalah penyeragaman bentuk huruf atau mengubah semua karakter huruf besar *(uppercase)* menjadi karakter huruf kecil *(lowercase)* pada dokumen teks.
2. *Remove* URL adalah penghapusan *link* URL (*Uniform Resource Locator*) yang terdapat pada teks. *Tweet* yang mengandung kata "http://" merupakan *tweet* dengan *link* URL.
3. *Unescape* HTML adalah penghapusan HTML *(HyperText Markup Language)* yang menggunakan tanda-tanda tertentu dan menghapus jejak karakter yang bisa dianggap sebagai *markup*.
4. *Remove mention* adalah penghapusan *mention* (penyebutan pengguna lain) pada data *tweet* perlu dilakukan karena umumnya tidak mengandung informasi yang penting.
5. *Remove number* adalah penghapusan semua angka yang terdapat pada teks.
6. *Remove punctuation* adalah penghapusan semua tanda baca selain alphabet pada teks.

2.4 *Feature Selection*

Feature selection merupakan tahapan untuk mengurangi dimensi dari data tekstual sehingga hasil dari data teks memiliki kualitas yang lebih baik[9]. Langkah-langkah yang dilakukan antara lain:

1. *Stemming* adalah proses mentrasformasikan kata-kata dalam dokumen menjadi kata akarnya (*root word*) atau kata dasar[1].
2. *Stopword* adalah sebagian kata dalam suatu korpus yang muncul dalam jumlah besar dan dianggap tidak penting atau tidak menggambarkan isi dari sebuah kalimat. Contoh dari *stopwords* adalah kata "yang", "dan", "dari", "ke", "ini", dan sebagainya. *Stopword* dihilangkan untuk mempercepat proses pengolahan data teks dan mendapatkan informasi yang relevan[16].

3. *Tokenizing* merupakan proses pemotongan teks berdasarkan tiap kata yang menyusunnya berdasarkan karakter spasi[11].

2.5 Pembobotan Data

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menghitung bobot setiap kata yang telah dianalisis. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model *term frequency* (TF) dan *inverse document frequency* (IDF). Tahap TF merupakan cara menentukan bobot setiap kata (*term*) pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Selanjutnya tahap IDF yaitu pengurangan dominasi kata yang sering muncul di berbagai dokumen. TF-IDF dihitung menggunakan persamaan sebagai berikut [12]:

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

Keterangan:

$W_{j,i}$: Pembobotan TF-IDF untuk *term* ke j pada dokumen ke i.
$n_{j,i}$: Jumlah kemunculan *term* ke j pada dokumen ke i.
$\sum_k n_{k,i}$: Jumlah kemunculan seluruh *term* pada dokumen ke i.
D : Banyaknya dokumen yang dibangkitkan.
d_j : Banyaknya dokumen yang mengandung *term* ke j.

2.6 *Silhouette Coefficient*

Silhouette coefficient merupakan salah satu metode yang digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode *cohesion* dan *separation* [6].

Dalam perhitungan nilai *silhouette coefficient*, terdapat komponen yaitu a_i dan b_i. Komponen a_i adalah rata-rata jarak data ke-i terhadap semua data lainnya dalam satu *cluster*, sedangkan komponen b_i adalah hasil perhitungan rata-rata jarak data ke-i terhadap semua data lainnya yang tidak dalam satu *cluster*, kemudian diambil nilai terkecil. Nilai *silhouette coefficient* berada pada rentang -1 hingga 1. Semakin besar nilai *silhouette coefficient* menunjukkan bahwa *cluster* tersebut merupakan *cluster* terbaik [6].

Berdasarkan penelitian [6], tahapan perhitungan *silhouette coefficient* adalah sebagai berikut :

1. Hitung rata-rata jarak dari suatu data misalkan i dengan semua data lain yang berada dalam satu *cluster* (a_i).
$$a_i^j = \frac{1}{m_j - 1} \sum_{r=1}^{m_j} d(x_i^j, x_r^j) \quad (2)$$
2. Hitung rata-rata jarak dari data *i* tersebut dengan semua data di *cluster* lain
$$d_i^j = \frac{1}{m_n} \sum_{r=1}^{m_j} d(x_i^j, x_r^{mj}) \quad (3)$$
3. Pengambilan nilai terkecil dari d_i^j sebagai nilai b_i^j
$$b_i^j = min\ d_i^j \quad (4)$$
4. Penghitungan nilai s*ilhouette coefficient* data
$$S_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (5)$$
5. Nilai *silhouette coefficient* dari sebuah *cluster*
$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} S_i^j \quad (6)$$
6. Nilai *silhouette coefficient global* ditentukan dengan menghitung rata-rata nilai *silhouette coefficient* semua *cluster* dengan rumus :
$$S_g = \frac{1}{k} \sum_{j=1}^{k} S_j \quad (7)$$

Keterangan :

i : indeks data
j : *cluster*
m_j : banyaknya data dalam *cluster j*
m_n : banyaknya data dalam satu *cluster*
$d(x_i^j, x_r^{mj})$: jarak data *i* pada *cluster j* dengan data *r* dalam *cluster* m_j
a_i^j : rata-rata jarak data ke-*i* terhadap semua data lainnya dalam satu *cluster j*
d_i^j : rata-rata jarak data ke-*i* terhadap semua data lainnya dalam satu *cluster j*
b_i^j : nilai minimum rata-rata jarak data *i* pada *cluster j* terhadap semua data di *cluster* lain
S_i^j : nilai *silhouette coefficient* data *i* pada *cluster j*
S_j : nilai *silhouette coefficient cluster j*
S_g : nilai *silhouette coefficient global*

2.6 *K-Means Clustering*

Analisis *cluster* digunakan untuk mengklasifikasi obyek atau kasus (responden) ke dalam kelompok yang relatif homogen yang disebut *cluster*, obyek atau kasus dalam setiap kelompok cenderung mirip satu sama lain dan tidak sama dengan obyek dari *cluster* lainnya [13]. *K-Means* merupakan salah satu metode analisis *cluster* yang melakukan partisi set data ke dalam sejumlah K *cluster* yang sudah ditetapkan di awal[18].

Proses clustering dimulai dengan mengidentifikasi data yang akan dicluster, x_{ij} (i=1,...,n; j=1,...,m) dengan n adalah banyaknya data yang akan dicluster dan m adalah banyaknya variabel. Pada awal iterasi, pusat setiap cluster ditetapkan secara bebas, c_{kj} (k=1,...,s; j=1,...,m) dengan k adalah banyaknya cluster. Kemudian dihitung jarak antara setiap data dengan setiap pusat cluster. Untuk melakukan penghitungan jarak data ke-i (xi) pada pusat cluster ke-k (ck), diberi nama D(x,c), dapat digunakan formula Euclidean, seperti persamaan (8) yaitu :

$$D(x,c) = d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj})^2} \quad (8)$$

Suatu data akan menjadi anggota dari cluster ke-k apabila jarak data tersebut ke pusat cluster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lainnya. Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap cluster.

Nilai pusat cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada setiap cluster tersebut, dengan menggunakan rumus pada persamaan (9) sebagai berikut :

$$\overline{c_{kj}} = \frac{\sum_{i=1}^{n} x_{ij}^k}{p_{kj}} \quad (9)$$

dengan $x_{ij}^k \in$ *cluster* ke-k

p_{kj} : banyaknya anggota *cluster* ke-k dari variabel j (tidak bernilai 0)

$\overline{c_{kj}}$: nilai pusat *cluster* yang baru

Cluster yang baik adalah *cluster* yang mempunyai homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*) dan heterogenitas yang tinggi antar *cluster* yang satu dengan *cluster* lainnya (*between cluster*) [10].

2.7 *Wordcloud*

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[2]. Konsep pembuatan *word cloud* yaitu kata-kata yang memiliki ukuran huruf terbesar merupakan kata yang paling penting dan banyak[14].

3. METODE PENELITIAN

3.1 Jenis dan Sumber Data

Penelitian ini menggunakan data kualitatif berupa 1500 *tweet* dari *timeline* akun Twitter @bliblidotcom. Data yang diambil merupakan data *tweet* tanpa *replies*. Pengambilan data *tweet* diperoleh dari API *(Application Programming Interface*) Twitter yang dilakukan pada tanggal 1 Maret 2020. Data yang diperoleh adalah data pada periode 1 Oktober 2019 sampai 1 Maret 2020.

3.2 Teknik Pengolahan Data

Analisis data pada penelitian ini menggunakan metode *text mining* dan *K-means clustering* dengan bantuan *software* R i386 3.6.2 dan Microsoft Excel 2013. Adapun langkah-langkah analisis data yang dilakukan adalah sebagai berikut:

1. *Extracting tweets*, pengambilan 1500 *tweets* terbaru di media sosial Twitter @bliblidotcom dengan ketentuan *tweets* tanpa *replies.* Data diakses melalui Twitter *Application Programming Interface* (API).
2. *Text preprocessing*, data teks yang telah diambil diolah melalui beberapa tahap, yaitu *case folding, remove* URL, *unescape* HTML, *remove mention, remove number*, dan *remove punctuation.*
3. *Feature selection*, tahapan untuk mengurangi dimensi dari sebuah data teks sehingga hasil memiliki kualitas yang lebih baik. Proses yang dilakukan adalah *stemming, stopword removal* dan *tokenizing*.
4. Pembobotan data teks dengan TF-IDF dalam bentuk *term-document matrix*
5. Penentuan jumlah *cluster* terbaik yang dilakukan berdasarkan hasil perhitungan nilai *Silhouette Coefficient*
6. Proses *clustering* data dengan metode *K-Means Clustering*
7. Analisis jenis konten *tweets* pada tiap *cluster*
8. Interpretasi data dengan *word cloud*

4. ANALISIS DAN PEMBAHASAN

4.1 *Extracting Tweets*

Extracting tweets bertujuan untuk mengumpulan data teks dari aplikasi Twitter dengan menggunakan Twitter API. Untuk melakukan *extracting tweets* dibutuhkan empat kode akses, yaitu API *Key*, API *Secret, Access Token, and Access Token Secret.* Kode-kode tersebut diperoleh setelah mendaftarkan akun Twitter pada https://developer.Twitter.com/en/apps. Pengambilan 5.000 tweets di akun Twitter @bliblidotcom dengan ketentuan *tweets* tanpa *replies*. Hasil *tweets* yang diperoleh pada ekstraksi *tweets* berjumlah 1846 *tweets* yang kemudian diambil 1500 *tweets* terbaru untuk dilakukan proses *clustering* data.

4.2 *Pre-Processing* Data

Proses ini merupakan tahap pengambilan data dan ekstraksi data yang mengubah format asli dan tidak terstruktur menjadi terstruktur agar dapat diolah untuk tahapan berikutnya. Tahapan yang dilakukan diantaranya adalah sebagai berikut:

1. *Case Folding*
   Case folding adalah tahapan mengubah semua huruf besar atau kapital pada data *Twitter* menjadi huruf kecil semua (*lowercase*) menggunakan fungsi ‘bli.corpus <- tm_map(bli.corpus,content_transformer(tolower))’.
2. *Remove* URL
   Remove URL akan menghapus *link* URL (*Uniform Resource Locator*) yang terdapat pada data *Twitter*. *Link* URL biasanya mengandung kata “http://”. *Remove* URL dilakukan dengan menggunakan fungsi ‘removeURL <- function(x) gsub("http[^[:space:]]*", "", x)’.

3. *Unescape* HTML
   Unescape HTML dilakukan untuk menghapus *file* HTML (*Hyper Text Markup Language*) yang menggunakan tanda-tanda tertentu dan menghapus jejak karakter yang bisa dianggap sebagai *markup*. *Unescape* HTML dilakukan dengan menggunakan fungsi 'unescapeHTML <- function(x){gsub("[^\x01-\x7F]", "", x)}'.
4. *Remove Mention*
   Remove mention dilakukan untuk menghilangkan kata yang mengandung "@" yang berarti menyebutkan *usename* pengguna Twitter lain dengan menggunakan fungsi 'removeMention <- function(x){gsub("@\\w+", "", x)}'.
5. *Remove Number*
   Semua angka yang terdapat pada dokumen teks akan dihapus dengan menggunakan fungsi 'bli.corpus <- tm_map(bli.corpus, toSpace, "[[:digit:]]")'.
6. *Remove Punctuation*
   Remove punctuation akan menghapus tanda baca yang ada pada data Twitter. Karena penelitian ini hanya mengklasifikasikan data teks, maka selain karakter alphabet akan dihapus dari data Twitter. *Remove punctuation* dilakukan dengan menggunakan fungsi 'bli.corpus <- tm_map(bli10s.corpus, toSpace, "[[:punct:]]")'.

Tabel 1. Hasil Proses *Preprocessing* Data

Tweet ke-	Sebelum *Preprocessing* Data	Sesudah *Preprocessing* Data
1	Heyho! Masuk di bulan Maret, Blibli langsung kasi promo buat kamu semua yang suka main games! #KarenaKamuNo1, diskon sampai 65%! Sikats~	heyho masuk di bulan maret blibli langsung kasi promo buat kamu semua yang suka main games karenakamuno diskon sampai sikats
2	Februari segera berakhir. Siap menyongsong bulan Maret, friends? <U+0001F4AA>	februari segera berakhir siap menyongsong bulan maret friends
3	belanja di lebih mudah dan hemat pakai gopay khusus sampai tanggal maret blibli fr	belanja di lebih mudah dan hemat pakai gopay khusus sampai tanggal maret blibli fr

4.3 *Feature Selection*

Tahapan ini dilakukan untuk mengurangi dimensi dari sebuah data Twitter dengan menghapus kata-kata yang tidak relevan sehingga proses pengelompokan lebih efektif dan akurat.

1. *Stemming*
   Stemming diperlukan untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen dengan cara menghilangkan imbuhan kata yang ada.
2. *Stopword*
   Remove stopwords digunakan untuk menghilangkan kata-kata dalam suatu korpus yang muncul dan dianggap tidak menggambarkan isi dari sebuah kalimat. Pemilihan kata yang bermakna dengan menghilangkan kata yang kurang penting dalam membangun model dapat meningkatkan hasil akurasi sistem klasifikasi. *Stopwords* yang digunakan pada penelitian ini berjumlah 773 kata yang diperoleh dari http://www.ranks.nl/stopwords/indonesian. Serta tambahan *stopword* sebanyak 1155 kata secara manual dengan menambahkan kata-kata yang tidak penting. Hasil *stopword* memiliki beberapa *tweet* yang hanya menyisakan kata-kata yang secara keseluruhan tidak mempunyai makna tertentu. *Tweets* ini akan dihilangkan pada tahap pembobotan TF.
3. Tokenizing
   Proses *tokenizing* dilakukan untuk memotong teks tiap kata berdasarkan spasi. Tabel 2 merupakan contoh proses *tokenizing* yang terbentuk setelah dilakukan *filtering stopwords.*

Tabel 2. Proses *Tokenizing*

No	Sebelum *Tokenizing*	Sesudah *Tokenizing*
6	sehat soyjoy harga spesial soylution active package	sehat soyjoy harga special soylution active package

4.5 Pembobotan Data

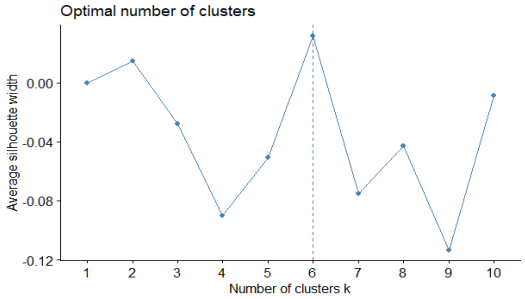
Pembobotan kata yang digunakan pada penelitian ini adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). Pembobotan data akan digunakan untuk membangun model klasifikasi. Data yang digunakan sejumlah 962 data yang telah melalui tahap penghapusan *tweet* yang hanya memiliki 1-3 kata dan tidak memiliki makna.

Tabel 3. Pembobotan dengan *Term Frequency Inverse Document Frequency*

	abadi	**absen**	**acara**	**accessories**	**Acer**	**acmic**	**action**	**...**	**zoom**
Tweet ke-1	0	0	0	0	0	0	0	...	0
Tweet ke-3	0	0	0	0	0	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
Tweet ke-1499	0	0	0	0	0	0	0	...	0

4.6 *Silhouette Coefficient*

Penentuan jumlah *cluster* terbaik yang dilakukan dalam penelitian ini menggunakan metode *Silhouette Coefficient* dengan menggunakan fungsi ‘fviz_nbclust (matriks_kata, kmeans, method = "*silhouette*", k.max=10)’. Perhitungan nilai *Silhouette Coefficient* dilakukan untuk data matriks_kata yang merupakan data pembobotan TF-IDF dengan k = 1 sampai k = 10. Hasil *Silhouette Coefficient* ditampilkan pada Gambar 1. Pada Gambar 1 terlihat bahwa titik tertinggi terdapat pada k=6 yang berarti jumlah *cluster* terbaik yaitu 6. Hasil perhitungan k terbaik digunakan untuk proses *clustering* dengan *K-means.*



Gambar 1. Nilai *Silhouette Coefficient* pada Jumlah *Cluster* 1 sampai 10

4.7 Analisis Jenis Konten *Tweet*

Penentuan jenis konten dari masing-masing *cluster tweet* dilakukan dengan menganalisis kata yang paling sering muncul pada masing-masing *cluster.* Kata-kata tersebut akan diambil 10 kata teratas yang paling sering muncul untuk menggambarkan jenis konten *tweets* pada setiap *cluster.* Fungsi yang digunakan untuk mengurutkan kata-kata tersebut yaitu ‘p <- sort(colSums(q),decreasing=TRUE)’.

Tabel 4. Hasil *Document Term Matrix* Pembobotan TF IDF

Cluster	Kata Paling Sering Muncul	Jenis Konten *Tweet*
1	diskon, produk, histeriasyok, harga, spesial, flash, sale, game, voucher, blogbliblifriends	penawaran diskon flash sale untuk voucher game dan produk lainnya
2	jalan, mall, tiket, voucher, weekend, cashback, diskon, gratis, handphone, histeriasyok	penawaran-penawaran pada sebuah mall waktu weekend seperti voucher, cashback, dan diskon
3	teman, final, pialapresidenesports, regional, kualifikasi, blibliesports, grand, barat, tanding, timur	Turnamen Blibli Esport Championship Indonesia
4	earphone, wireless, beats, kabel, kualitas, aksesoris, canggih, charger, denger, lagu	penawaran barang elektronik aksesoris handphone seperti charger dan earphone
5	libur, foto, kamera, diskon, visa, tiket, game, hasil, lensa, produk	penawaran produk yang digemari saat liburan seperti kamera, tiket, game dan produk lainnya
6	pasangan, kado, valentine, lagu, diskon, orang, romantis, sayang, bangun, bucin	penawaran diskon kado valentine untuk pasangan

Gambar 2, Gambar 3, Gambar 4, Gambar 5, Gambar 6, dan Gambar 7 menunjukkan visualisasi dari jenis konten *tweets* pada *cluster* 1, *cluster* 2, *cluster* 3, *cluster* 4, *cluster* 5, dan *cluster* 6.

Gambar 2. *Word Cloud Cluster* 1

Gambar 3. *Word Cloud Cluster* 2

Gambar 4. *Word Cloud Cluster* 3

Gambar 5. *Word Cloud Cluster* 4

Gambar 6. *Word Cloud Cluster* 5

Gambar 7. *Word Cloud Cluster* 1

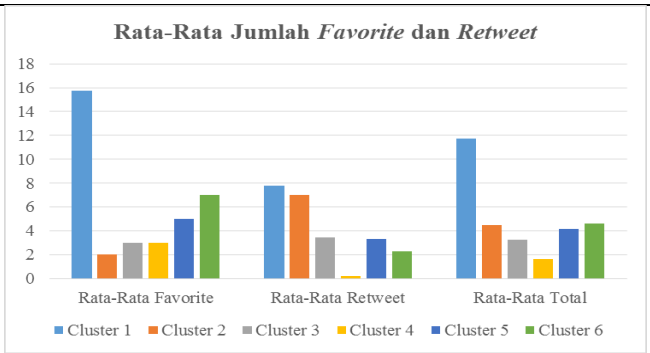
4.8 Analisis Konten yang Disukai *Followers* Twitter @bliblidotcom

Penentuan konten yang disukai oleh *followers* akun Twitter @bliblidotcom dapat dilakukan berdasarkan perhitungan rata-rata jumlah *favorite* dan *retweet* masing- masing *cluster*. Perhitungan rata-rata jumlah *favorite* dan *retweet* untuk masing-masing cluster terlihat pada Tabel 5.

Pada Gambar 8 terlihat bahwa *cluster* yang memiliki rata-rata *favorite* dan *retweet* tertinggi adalah *cluster* 1 yaitu *tweet* mengenai penawaran diskon *flash sale* untuk *voucher game* dan produk lainnya. Sedangkan *cluster* yang memiliki rata-rata *favorite* dan *retweet* terendah adalah *cluster* 1 yaitu *tweet* mengenai penawaran barang elektronik aksesoris *handphone* seperti *charger* dan *earphone*. Blibli Indonesia diharapkan mampu memanfaatkan konten diskon dan *flash sale* sebagai sarana periklanan semaksimal mungkin.

Tabel 5. Hasil Perhitungan Rata-Rata *Favorite* dan *Retweet*

Cluster	Jumlah Favorite	Jumlah Retweet	Jumlah Tweet	Rata-Rata Favorite	Rata-Rata Retweet	Rata-Rata Total
1	13005	6411	827	15.726	7.752	11.739
2	30	105	15	2	7	4.500
3	123	142	41	3	3.463	3.232
4	15	1	5	3	0.200	1.600
5	230	152	46	5	3.304	4.152
6	196	63	28	7	2.250	4.625



Gambar 8. Grafik Rata-Rata Jumlah *Favorite* dan *Retweet*

Masing-masing *cluster* memiliki *tweet* dengan jumlah *favorite* dan *retweet* paling banyak diantara *tweets* lain pada *cluster* tersebut. *Tweet* tersebut juga memiliki periodenya masing-masing.

Tabel 6. *Tweet* dengan Jumlah *Favorite* dan *Retweet* Paling Banyak pada Tiap *Cluster*

Cluster	*Tweet* ke-	Isi *Tweet*	Jumlah *Favorite* *Retweet*	Tanggal
1	944	Mimin gak nyangka bisa sedeket ini sama EXO. Rasanya Mimin seperti sedang bermimpi <U+0001F634> #EXplOrationinJKT… https://t.co/pH41CiGPSC	4260	23/11/2019
2	917	Daripada memberi amarah ke padatnya jalanan, mendingan mimin bagi-bagi voucher belanja. Pasti mau kan kamu? Coba b… https://t.co/niI9MXiEYH	143	26/11/2019
3	306	Holaaaaa! Selamat siang pejuang Esports! Sekarang aku lagi di Grand Final #PialaPresidenEsports2020 yang berlangsu… https://t.co/TKophjIzhF	19	01/02/2020
4	673	Lagi nge-trend banget nih earphone wireless dan kualitasnya ajib! Nah, ini nih alasan kenapa earphone wireless dig… https://t.co/kCFyVc1hBe	5	19/12/2019
5	157	Sudah cek jadwal libur lebaran? Segerakan membeli tiket kereta ke kotamu, agar bisa berkumpul bersama orang tua dan… https://t.co/1NLn8GSjga	17	16/02/2020
6	121	Gerimis-gerimis, dengerin lagu iKon. Hari Kamis, harinya diskon! #KarenaKamuNo1 diskon sampai 50% buat yang lagi… https://t.co/4S4i4pjUbi	159	20/02/2020

5. KESIMPULAN

Cluster tweet yang terbentuk adalah konten mengenai aktivitas perbelanjaan, penawaran-penawaran pada sebuah *mall,* Turnamen Blibli Esport Championship Indonesia, penawaran barang elektronik aksesoris *handphone,* aktivitas rekreasi, serta penawaran diskon kado *valentine*.

Jumlah anggota masing-masing *cluster* kurang merata. *Cluster* 1 memiliki anggota sebanyak 827, c*luster* 2 sebanyak 15, c*luster* 3 sebanyak 41, c*luster* 4 sebanyak 5, c*luster* 5 sebanyak 46, dan c*luster* 6 sebanyak 28. *Cluster* 1 memiliki paling banyak anggota daripada *cluster* yang lain karena perbedaan bobot yang cukup jauh dan banyaknya kata yang sama terdapat pada *cluster* 1 seperti kata "diskon" dan "promo" yang jumlahnya sangat banyak. Sebagian besar *tweet* yang ditulis oleh akun @bliblidotcom juga memuat tentang aktivitas perbelanjaan yang termasuk dalam *cluster* 1.

Setiap *cluster* memiliki periodenya masing-masing sesuai dengan isi *tweet* yang disampaikan oleh @bliblidotcom. *Cluster* 1 banyak terdapat pada periode bulan Desember 2019, *cluster* 2 banyak terdapat pada periode akhir bulan, *cluster* 3 berlangsung dari November 2019 sampai Februari 2020, *cluster* 4 banyak terdapat pada Desember 2019, *cluster* 5 banyak terdapat pada libur akhir tahun Desember 2019, serta *cluster* 6 banyak terdapat pada periode Februari 2020.

Strategi yang perlu dilakukan untuk mengolah *term* kata tersebut agar jumlah anggota *cluster* lebih seimbang adalah dengan cara pemilihan kata pada proses *stemming* dan *stopwords.* Kata-kata yang diolah dalam proses tersebut harus lebih teliti. Pengelompokkan nama-nama merk produk menjadi suatu kategori juga diperlukan agar tidak mempengaruhi perbedaan bobot yang terlalu jauh.

Dari pola yang diperoleh dapat diberikan beberapa saran:

1. Jenis konten dengan rata-rata jumlah *retweet* dan *favorite* tertinggi yaitu mengenai aktivitas perbelanjaan seperti diskon produk dan flash sale. Blibli dapat memaksimalkan konten berupa diskon dan flash sale untuk lebih menarik minat pembeli dan lebih banyak mendapatkan respon positif.
2. Rata-rata terendah jumlah *retweet* dan *favorite* yaitu mengenai penawaran barang elektronik aksesoris *handphone*, Blibli dapat menggunakan konten penawaran barang elektronik ini dengan tambahan diskon atau promo-promo produk yang lebih menarik.
3. Jumlah *retweet* dan *favorite* Blibli memiliki nilai rata-rata yang rendah, Blibli dapat meningkatkan jumlahnya dengan menggunakan konten seperti konten kuis berhadiah dan *giveaway* yang mengharuskan *followers* menyebarkan *tweet* tersebut kepada pengguna lain sebagai syarat mengikuti kuis berhadiah tersebut.

Beberapa perbaikan yang dapat dilakukan untuk penelitian selanjutnya adalah menggunakan metode *clustering* yang lebih efisien serta memiliki kecocokan dengan data yang relatif *homogen*. *Output* yang lebih akurat dapat diperoleh dengan menggunakan *composite tokenization* pada proses *tokenizing*. Pada tahap *preprocessing* merupakan tahap yang paling penting dan riskan. Perhatikan pemilihan kata pada tahap *stopwords* dan penggunaan sintaks diperlukan agar mempermudah proses *stemming* dan mendapatkan hasil yang lebih akurat dan teliti.

DAFTAR PUSTAKA

[1] Afuan, L. 2013. Stemming Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Porter. Jurnal Telematika, Vol. 6 No. 2.

[2] 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.

[3] Feldman, R dan Sanger, J. 2007. The Text Mining Handbook. New York: Cambridge University Press.
[4] Go, A., Bhayani, R., dan Huang, L. 2009. Twitter Sentiment Classification using Distant Supervision. Stanford: Stanford University.
[5] Gupta, V dan Lehal, G. S. 2009. A Survey of Text Mining Techniques and Applications. Jurnal Emerging Technologies in Web Intelligence Vol.1, No.1: Hal 60-7.
[6] Handoyo, R., Mangkudjaja, R., & Nasution, S. M. 2014. Perbandingan Metode Clustering menggunakan Metode Single Linkage dan K-means pada Pengelompokan Dokumen. Jurnal Sifo Mikroskil, Vol. 15, No.2, Hal: 73-82.
[7] Harjanta, A. J. T. 2015. Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining. Jurnal Informatika UPGRIS Vol. 1.
[8] Hootsuite. 2020. Local Insights. https://datareportal.com/reports/digital-2020-indonesia. Diakses: 13 April 2020.
[9] Indraloka, D. S. dan Santosa, B. 2017. Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia. Jurnal Sains dan Seni ITS, Vol. 6, No.2: 2337-3520.
[10] Laeli, S. 2014. Analisis Cluster dengan Average Linkage Method and Ward's Method untuk Data Responden Nasabah Asuransi Jiwa Unit Link. Yogyakarta: Universitas Negeri Yogyakarta.
[11] Nurhuda, F., Sihwi, S. W., dan Doewes, A. 2013. Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. Jurnal IT SMART, Vol. 2, No. 2, Hal: 35-42.
[12] Salton, G. dan Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Jurnal Information Processing and Management Vol.24, No. 5, Hal: 512-523.
[13] Supranto, 2004. Analisis Multivariat Arti dan Interpretasi. Jakarta: PT. Rineka Cipta.
[14] Tessem, B., Bjornestad, S., Chen, W. & Nyre, L., 2015. Word Cloud Visualization of Locative Information. Journal of Location Based Services, Hal. 254-272.
[15] Twitter. 2020. Tentang Twitter. www.twitter.com. Diakses: 14 April 2020.
[16] Utomo, M. S. 2015. Stopword Dinamis dengan Pendekatan Statistik. Jurnal Informatika Upgris, Vol. 1, No. 2.
[17] Wahid, D. H. dan Azhari. 2016. Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity. Indonesian Journal of Computing and Cybernetics Systems (IJCCS), Vol. 10 , No. 2, Hal: 207-218.
[18] Wu, X. dan Kumar, V. 2009. The Top Ten Algorithms in Data Mining. USA: Chapman and Hall/CRC.