Text Clustering pada Akun *TWITTER* Layanan Ekspedisi JNE, J&T, dan Pos Indonesia Menggunakan Metode *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) dan *K-Means*

Devi Putri Isnarwaty dan Irhamah
Departemen Statistika, Fakultas Matematika, Komputasi dan Sains Data
Institut Teknologi Sepuluh Nopember (ITS)
e-mail: irhamah@statistika.its.ac.id[2]

Abstrak*—Tingginya minat masyarakat untuk berbelanja *online* membuat meningkatnya layanan ekspedisi yang digunakan untuk mengirimkan produk dari transaksi secara *online* maupun *offline*. Ada banyak perusahaan ekspedisi yang populer di Indonesia misalnya JNE, J&T, dan Pos Indonesia. Perusahaan ekspedisi gencar melakukan promosi lewat media sosial, misalnya saja Twitter. Akun Twitter ini dapat digunakan sebagai media bagi pelanggan untuk memberikan pendapat, kritik maupun saran, dan bagi pihak perusahaan untuk memberikan tanggapan maupun informasi. Analisis terhadap twitter yang dikirim, ber-guna bagi perusahaan untuk meningkatkan performa layanan. Dokumen twitter berupa teks sehingga diperlukan text mining untuk menganalisisnya. Dalam penelitian ini, text clustering di-gunakan untuk mengelompokkan pendapat menjadi beberapa kategori. Metode yang digunakan adalah metode *K-Means* dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). DBSCAN adalah sebuah metode yang membentuk *cluster* dari data-data yang saling berdekatan/rapat, sedangkan data yang saling berjauhan tidak akan menjadi anggota cluster. Sedangkan *K-Means* merupakan teknik clustering yang sederhana dan cepat dalam proses clustering obyek serta mampu mengelompokkan data dalam jumlah yang cukup besar. Ber-dasarkan nilai *silhouette coefficient,* metode DBSCAN lebih baik daripada K-Means dalam mengelompokkan *tweet* yang ditujukan kepada layanan ekspedisi JNE, J&T, dan Pos Indonesia karena menghasilkan *silhouette coefficient* yang lebih tinggi.

Kata Kunci— Clustering, Ekspedisi, DBSCAN, K-Means, Text Mining.

I. PENDAHULUAN

PERKEMBANGAN belanja *online* di masyarakat Indonesia ber-kembang sangat pesat dalam waktu yang relatif singkat. Kegiatan berbelanja secara *online* didorong oleh per-tumbuhan industri *e-commerce* di Indonesia yang cukup tinggi. Tingginya minat masyarakat untuk berbelanja *online* dan pengiriman barang, tentunya tidak lepas dari dibutuhkannya terhadap layanan pengiriman atau ekspedisi yang digunakan untuk mengirimkan barang/produk dari transaksi jual/beli secara *online* hingga berada di tangan *customer*. Hal ini menjadikan peluang bisnis bagi perusahaan ekspedisi.

Ada banyak perusahaan pengiriman barang yang populer di Indonesia misalnya saja JNE, J&T, dan Pos Indonesia. Perusahaan ekspedisi gencar melakukan promosi lewat media sosial. Bukan hanya promosi tetapi juga berinteraksi antara perusahaan ekspedisi dengan pelanggan. Perusahaan ekspedisi dapat menerima pendapat secara terbuka yang diberikan oleh pelanggan melalui media sosial.

Setiap perusahaan, memiliki customer service yang bertugas untuk memberikan pelayanan kepada pelanggan ter-masuk menerima keluhan/masalah yang sedang dihadapi oleh pelanggan tersebut. *Customer service* bukan hanya dapat di-akses melalui telepon atau email tetapi juga dapat diakses melalui media sosial Twitter. Perusahaan ekspedisi JNE, J&T, dan Pos Indonesia memiliki akun Twitter *customer care* yaitu @JNECare untuk JNE, @jntexpressid untuk J&T dan @PosIndonesia untuk Pos Indonesia. Akun tersebut digunakan sebagai layanan pelanggan secara *online* yang disediakan untuk memberikan pendapat, kritik, saran ataupun keluhan dari pelanggan. Komentar dalam *Twitter* berbentuk teks, sehingga perlu dilakukan analisis *text mining*. *Text mining* dapat memberikan solusi dari permasalahan seperti pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar. Salah satu teknik analisis dalam *text mining* adalah *text clustering*.

Pendapat masyarakat yang ditujukan pada akun media sosial *customer care* layanan ekspedisi dapat dikelompokkan menjadi beberapa kategori atau *cluster*. Penentuan kategori dari *tweet* yang ditujukan kepada layanan ekspedisi JNE, J&T, dan Pos Indonesia dalam mengantisipasi banyaknya tweet, seperti membuat template tanggapan untuk setiap kategori, dan mengelompokkan tweet yang masuk berdasarkan kategori dari hasil *cluster*. Pada penelitian ini, dilakukan dengan meng-gunakan metode *K-Means* dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). DBSCAN adalah algoritma pengelompokkan yang didasarkan pada ke-padatan (*density*) data [1]. Metode ini membentuk *cluster* dari data-data yang saling berdekatan, sedangkan data yang saling berjauhan tidak akan menjadi anggota cluster [2]. Sedangkan *K-Means* adalah metode *clustering* yang memiliki kemampuan untuk mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi. *K-Means* adalah teknik yang cukup sederhana dan cepat dalam proses *clustering* obyek [3]. Hasil evaluasi kedua metode tersebut menggunakan nilai *silhouette coefficient*. Metode terbaik dipilih berdasarkan nilai *silhouette coefficient* yang tinggi. Penelitian ini diharap-kan dapat membantu pihak layanan ekspedisi untuk me-ngetahui kategori pendapat yang paling sering diberikan oleh masyarakat.

II. TINJAUAN PUSTAKA

A. Text Mining

Text mining dapat didefinisikan sebagai proses menggali informasi dimana pengguna bisa berinteraksi dengan beberapa sumber dari waktu ke waktu menggunakan *tools analysis*. *Text mining* bertujuan mengekstraksi informasi yang berguna dari beberapa sumber data melalui identifikasi dan eksplorasi pola yang menarik [4].

B. Text Preprocessing

Text Preprocessing merupakan sebuah langkah penting dalam Data Mining untuk membuat data mentah menjadi data yang berkualitas. Data perlu dilakukan *preprocessing* karena dalam data mentah terdapat data yang tidak lengkap, *noise*, dan tidak konsisten [5]. Langkah – langkah dalam praproses teks adalah sebagai berikut.

1. Data *Cleaning*, tahap untuk menghilangkan kata yang tidak diperlukan misalnya karakter HTML, link URL, *username* (@username), emoticons, dan hashtag (#) [6].
2. *Case Folding*, tahap untuk menghilangkan angka dan tanda baca, serta mengubah karakter teks menjadi huruf kecil semua [7].
3. *Stemming*, tahap untuk mendapatkan kata dasar. Sistem kerja tahap *stemming* ini adalah menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi dari awalan dan akhiran) [8].
4. *Stopwords*, tahap untuk menghilangkan kosakata yang bukan termasuk kata unik atau tidak menyampaikan pesan apapun secara signifikan pada teks. Kosakata yang dimaksud seperti kata penghubung dan kata keterangan misalnya “oleh”, “di”, “yang”, “jadi”, dan sebagainya [9].
5. Tokenizing, tahap untuk memutuskan kata per kata pada kalimat. Tahapan ini bertujuan untuk memecah yang semula berupa kalimat menjadi potongan-potongan kata [10].

C. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk mengukur seberapa penting suatu kata dalam dokumen. Frekuensi kemunculan setiap kata dalam dokumen diberikan sebagai skema pembobotan kata yang diasumsikan bahwa *stopwords* yang paling sering muncul dalam teks di-hapus terlebih dahulu. Kata-kata yang diberikan bobot (*weight*) dihitung frekuensi kemunculannya dalam sebuah dokumen [11]. Berikut adalah persamaan yang membentuk TF-IDF yang dapat dilihat pada persamaan (1) dan (2).

$$W_{i,j} = TF_{i,j} \times IDF_j , \quad (1)$$

$$IDF_j = \log\left(\frac{N}{DF_j}\right), \quad (2)$$

keterangan :

$W_{i,j}$ = bobot dari kata ke *j* pada *tweet* ke *i*
DF_j = banyaknya *tweet* yang mengandung kata *j*
$TF_{i,j}$ = jumlah kemunculan kata ke *j* pada *tweet* ke *i*
IDF_j = *inverse document frequency* pada kata ke *j*
N = jumlah keseluruhan *tweet*

D. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density Based Spatial Clustering Algorithm with Noise (DBSCAN) adalah algoritma pengelompokkan yang didasarkan pada kepadatan (*density*) data [1]. Algoritma DBSCAN membutuhkan dua parameter penting, yaitu parameter radius (*Eps*) dan jumlah minimum poin untuk membentuk kelompok (*MinPts*). Algoritma dari DBSCAN adalah sebagai berikut [12].

1. Menentukan parameter *MinPts* dan *Eps.*
2. Pilih *tweet p* secara acak
3. Menghitung jumlah *tweet* yang ditentukan oleh parameter radius (*Eps*). Jika jumlahnya mencukupi (lebih dari atau sama dengan ε), data akan ditandai sebagai inti (*core point*).
4. Menghitung jarak titik *core point* dengan *point* yang lain menggunakan jarak *Euclidean*. Berikut adalah rumus jarak *Euclidean* yang ditujukan pada persamaan (3).

$$d_{ip} = \sqrt{\sum_{j=1}^{m} (x_{ji} - y_{jp})^2} , \quad (3)$$

keterangan :

d_{ip} = jarak *Euclidean* dari *tweet* ke-*i* ke pusat *cluster* ke-*k*
x_{ji} = frekuensi kemunculan kata ke-*j* pada *tweet* ke-*i*
y_{jp} = frekuensi kemunculan kata ke-*j* pada titik pusat ke-*p*
m = banyak kata

5. Buat *cluster* baru dengan menambahkan *tweet p* ke dalam *cluster*.
6. Melakukan identifikasi pada data yang ditandai sebagai *core point*
7. Lanjutkan proses sampai semua point telah diproses.
8. Jika ada *tweet* yang tidak masuk ke dalam cluster manapun akan ditandai sebagai noise.

DBSCAN mencari *cluster* dengan memeriksa parameter radius (*Eps*) dari setiap titik dalam dataset. Jika *Eps* pada *tweet p* berisi lebih dari MinPts, sebuah cluster baru dengan *p* sebagai *core point* terbentuk [13].

E. K-Means

K-Means adalah teknik yang cukup sederhana dan cepat dalam proses *clustering* obyek (*clustering*) serta mampu mengelompokkan data dalam jumlah cukup besar. Berikut adalah algoritma dari metode K-Means [14].

1. Memilih secara acak *k centroid* awal dalam data
2. Menentukan jarak setiap kata terhadap pusat *cluster*
3. Mengelompokkan setiap kata berdasarkan kedekatannya dengan *centroid* (jarak terkecil) menggunakan *Euclidean distance*. Berikut adalah rumus jarak *Euclidean* yang ditujukan pada persamaan (4).

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ji} - y_{jk})^2} , \quad (4)$$

keterangan :

d_{ik} = jarak *Euclidean* dari *tweet* ke-*i* ke pusat *cluster* ke-*k*
x_{ji} = frekuensi kemunculan kata ke-*j* pada *tweet* ke-*i*
y_{jk} = frekuensi kemunculan kata ke-*j* pada pusat *cluster* ke-*k*
m = banyak kata

4. Hitung ulang pusat *cluster* (*centroid*) baru menggunakan

$$v_{ik} = \frac{1}{n_k} \sum_{j=1}^{m} x_{ji} , \quad (5)$$

keterangan :

v_{ik} = *centroid* (rata-rata *cluster* ke-*k* untuk *tweet* ke-*i*)
n_k = banyaknya *tweet* yang menjadi anggota *cluster* ke-*k*
x_{ji} = frekuensi kemunculan kata ke-*j* yang berada dalam *cluster* tersebut untuk *tweet* ke-*i*

5. Ulangi langkah 2 hingga 4, sampai anggota yang ada pada tiap *cluster* tidak berubah.

Jika jumlah *cluster k* belum diketahui, dapat menggunakan *Variance Ratio Criterion* (*VRC*) untuk menentukan jumlah *cluster k* yang optimum [15]. Berikut adalah rumus dari *Variance Ratio Criterion* yang dapat dilihat pada persamaan (6), (7), dan (8).

$$VRC = \frac{BGSS/(K-1)}{WGSS/(N-K)}, \quad (6)$$

$$BGSS = \sum_{k=1}^{K}\sum_{l=1,l\neq k}^{K}\sum_{j=1}^{m}(y_{jk}-y_{jl})^2, \quad (7)$$

$$WGSS = \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{m}(x_{ji}-y_{jk})^2, \quad (8)$$

keterangan :

K = banyak *cluster*
N = banyak *tweet*
m = banyak kata
I = banyak *tweet* yang menjadi anggota *cluster* ke-*k*
x_{ji} = frekuensi kemunculan kata ke-*j* yang berada dalam *cluster* tersebut untuk *tweet* ke-*i*
y_{jk} = frekuensi kemunculan kata ke-*j* pada pusat *cluster* ke-*k*
y_{jl} = frekuensi kemunculan kata ke-*j* pada pusat *cluster* ke-*l*

F. Silhouette Coefficient

Silhouette coefficient merupakan metode yang digunakan untuk mengevaluasi hasil *clustering* dengan memeriksa se-berapa baik kelompok-kelompok (*cluster*) yang dihasilkan [12]. Berikut adalah langkah-langkah perhitungan *silhouette coefficient*.

1. Menghitung jarak rata-rata dari tweet *i* dengan semua *tweet* yang ada di dalam *cluster* yang sama

$$a(i) = \frac{\sum_{j\in C_i, i\neq j} dist(i,j)}{|C_i|-1}, \quad (9)$$

keterangan :

C_i = banyak *tweet* dalam *cluster* C_i
j = *tweet* lain dalam *cluster* C_i
$dist(i, j)$ = jarak *Euclidean* antara *tweet i* dengan *tweet j*

2. Menghitung jarak rata-rata dari *tweet i* dengan semua *tweet* yang berada di *cluster* berbeda dan didapatkan nilai terkecil.

$$b(i) = \min_{C_l:1\leq l\leq k, l\neq 1}\left\{\frac{1}{|C_l|}\sum_{j\in C_l} dist(i,l)\right\}, \quad (10)$$

keterangan :

C_l = banyak *tweet* dalam *cluster* C_l
l = *tweet* lain dalam *cluster* yang berbeda
$dist(i, l)$ = jarak *Euclidean* antara *tweet i* dengan *tweet l*

3. Hitung *silhouette coefficient*

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \quad (11)$$

Nilai *silhouette coefficient* $s(i)$ adalah antara -1 dan 1. Ketika nilai $s(i)$ ada di sekitar 1 yang berarti memiliki pengelompokkan jauh dari *cluster* lain [16].

G. Word Cloud

Word cloud adalah visualisasi umum lainnya berdasarkan frekuensi. Di *word cloud*, kata-kata diwakili dengan ukuran *font* yang bervariasi. Dalam *word cloud* sederhana, hanya satu dimensi informasi yang ditampilkan dengan ukuran *font* yang sesuai dengan frekuensi *n* gram yang berarti bahwa semakin besar kata dalam *world cloud* dapat diubah sebagai informasi baru, misalnya dalam warna dan pengelompokkan [17].

H. Layanan Ekspedisi

Perusahaan Jasa Pengiriman atau ekspedisi merupakan sebuah perusahaan yang bergerak pada bidang layanan pengiriman, yang dalam hal ini adalah pengiriman barang [18].

III. METODE PENELITIAN

A. Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data yang diambil dari kumpulan *tweet* yang ditujukan kepada *customer care* dari Layanan Ekspedisi JNE (@JNECare), *customer care* dari Layanan Ekspedisi J&T(@jntexpressid), dan *customer care* dari Layanan Ekspedisi Pos Indonesia (@PosIndonesia) selama 4 Februari 2019 hingga 7 Mei 2019 dengan mengguna-kan Twitter API. Variabel penelitian yang digunakan adalah frekuensi kemunculan kata dasar dari setiap *tweet* yang telah dilakukan *preprocessing*, dinotasikan sebagai variabel *A* dengan skala rasio.

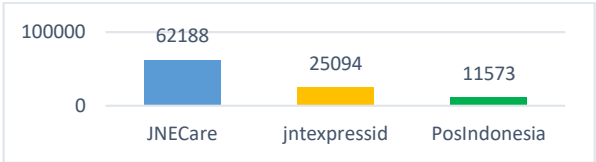
B. Langkah Analisis

Langkah analisis yang digunakan pada penelitian ini adalah sebagai berikut.

1. Mengambil data *tweet* dengan menggunakan *Twitter* API
   a. Memasukkan *keyword* yang berhubungan dengan akun *Twitter* @JNECare, @jntexpressid dan @PosIndonesia
   b. Menyimpan hasil *crawling* ke database
2. Melakukan *text preprocessing* pada data *tweet* layanan ekspedisi JNE, J&T, dan Pos Indonesia
3. Mengubah data *tweet* ke dalam bentuk frekuensi kemunculan kata menggunakan metode TF-IDF.
4. Melakukan *clustering* data.
   a. Melakukan *clustering* dengan metode DBSCAN.
      i. Melakukan *clustering* sesuai dengan algoritma dari DBSCAN.
      ii. Menghitung nilai *silhouette coefficient* setiap kombinasi *Eps*.
   b. Melakukan *clustering* dengan metode *K-Means*.
      i. Melakukan *clustering* sesuai dengan algoritma dari *K-Means*
   c. Memilih hasil *clustering* terbaik dengan melihat nilai *silhoutte coefficient* terbesar.

IV. ANALISIS DAN PEMBAHASAN

A. Karakteristik Data



Gambar 1. Perbandingan Jumlah *Tweet* pada Akun Layanan Ekspedisi.

Dari hasil *crawling* menggunakan Twitter API, didapatkan 62.188 *tweet* yang ditujukan kepada akun layanan ekspedisi JNE yang diambil dari tanggal 17 Februari 2019–7 Mei 2019, 25.094 *tweet* yang ditujukan kepada akun layanan ekspedisi J&T yang diambil dari tanggal 17 Februari 2019–7 Mei 2019, dan 11.573 *tweet* yang ditujukan kepada akun layanan ekspedisi Pos Indonesia yang diambil dari tanggal 4 Februari 2019–7 Mei 2019.

Berdasarkan Gambar 1 menunjukkan bahwa *tweet* yang paling banyak diberikan oleh masyarakat adalah *tweet* pada akun layanan ekspedisi JNE. Hal ini dapat dilihat bahwa layanan ekspedisi JNE lebih sering digunakan oleh masyarakat untuk mengirim barang.

Sebelum melakukan analisis *clustering* perlu dilakukan *text preprocessing* pada kumpulan *tweet* yang dihasilkan. *Text preprocessing* dilakukan agar data mentah yang didapatkan dapat menjadi tersruktur untuk mempermudah analisis.

B. *Layanan Ekspedisi JNE*

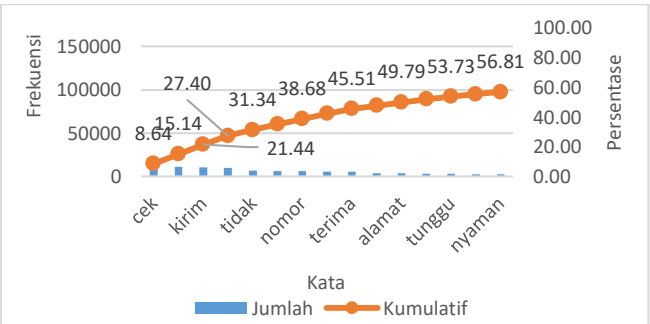
Berikut adalah struktur data berdasarkan frekuensi kemunculan kata yang didapatkan setelah dilakukan *text prepocessing* pada akun layanan ekspedisi JNE yang ditampilkan pada Tabel 2.

Tabel 2.
Struktur Data Layanan Ekspedisi JNE Setelah Dilakukan *Preprocessing*

Tweet ke-	Kata						
	alamat	…	cek	…	kirim	…	nomor
1	0	…	0	…	0	…	0
2	2	…	1	…	1	…	0
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
15.003	1	…	0	…	0	…	2
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
35.123	0	…	0	…	0	…	0

Tabel 2 menunjukkan bahwa frekuensi kemunculan setiap kata pada setiap *tweet* yang didapatkan dari proses *text pre-processing*. Diketahui bahwa jumlah kata telah berkurang dikarenakan ada beberapa kata yang tidak memiliki makna dalam kalimat dihapus dan tidak digunakan. Jumlah dari kata yang dihasilkan setelah dilakukan *text preprocessing* adalah 339 kata dan kata-kata tersebut merupakan variabel penelitian yang akan digunakan. Selanjutnya dilakukan visualisasi data menggunakan *word cloud* dengan teknik *n*-gram, yaitu dengan *n* sebesar 1 atau *unigram*. Berikut adalah *word cloud* yang dibentuk berdasarkan bobot TF-IDF dari setiap kata pada akun layanan JNE yang ditampilkan pada Gambar 2.

Gambar 2. Visualisasi *Word Cloud* pada Layanan Ekspedisi JNE.

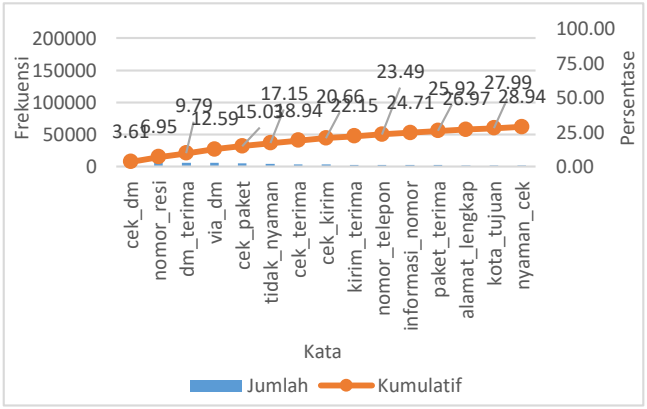


Gambar 3 Diagram Pareto *Unigram* Layanan Ekspedisi JNE.

Berdasarkan diagram pareto pada Gambar 3, diketahui bahwa jika diambil 30% dari total frekuensi kemunculan kata, maka kata yang sering muncul adalah pada *tweet* yang ditujukan pada layanan ekspedisi JNE adalah "cek", "paket", "kirim", dan "dm". Hal ini sesuai dengan hasil visualisasi *word cloud unigram* pada Gambar 2. Kata yang lebih diutamakan untuk ditindak lanjuti adalah 30% dari total frekuensi ke-munculan kata. Berikut adalah visualisasi *word cloud* meng-gunakan bigram yang ditampilkan pada Gambar 4.

Gambar 4. Visualisasi *Word Cloud Bigram* pada Layanan Ekspedisi JNE.

Gambar 4 adalah diagram pareto yang menunjukkan bahwa jika pasangan kata diambil 15% dari total frekuensi kemunculan kata, maka pasangan kata yang sering muncul adalah pada *tweet* yang ditujukan pada layanan ekspedisi JNE adalah "cek_dm", "nomor_resi", "dm_terima", dan "via_dm". Pasangan kata tersebut sesuai dengan hasil visualisasi *word cloud bigram* pada Gambar 3. Kata yang lebih diutamakan untuk ditindak lanjuti adalah 15% dari total frekuensi ke-munculan kata. Kata "cek dm" merupakan bentuk permintaan pelanggan kepada pihak ekspedisi JNE untuk membaca dan merespon dm atau pesan yang dikirimkan pelanggan. Pasangan kata "dm terima" adalah bentuk informasi yang disampaikan oleh pihak ekspedisi JNE bahwa dm yang dikirim oleh pelanggan telah diterima, sedangkan pasangan kata "nomor resi" menunjukkan alat bukti pelacakan barang atau paket yang dikirimkan.



Gambar 5. Diagram Pareto *Bigram* Layanan Ekspedisi JNE.

1. *Clustering* menggunakan Metode DBSCAN

Hasil *clustering* dengan metode DBSCAN dengan meng-gunakan *MinPts* sebesar 50 dengan berbagai kombinasi *Eps* memperoleh nilai *silhouette coefficient* untuk layanan ekspedisi JNE.

Tabel 4 menunjukkan bahwa nilai *silhouette coefficient* tertinggi dilakukan menggunakan parameter *Eps* sebesar 0,6. Nilai *silhouette* coefficient yang dihasilkan dengan

mengguna-kan parameter *Eps* sebesar 0,6 dan *MinPts* sebesar 50 adalah 0,26518. Jumlah *cluster* yang didapatkan dari *clustering* menggunakan metode DBSCAN dengan *MinPts* sebesar 50 dan *Eps* sebesar 0,6 adalah sebanyak 18 *cluster* dengan 15.209 *tweet* sebagai *noise*. Oleh karena itu, dengan menggunakan *MinPts* sebesar 50 didapatkan *Eps* optimum sebesar 0,6.

Tabel 4.
Nilai *Silhouette Coefficient* dari Metode DBSCAN

Eps	*Silhouette Coefficient*
0,6	0,26518
0,61	0,21345
0,62	0,19498
0,63	0,09340
0,64	0,12935
0,65	0,01673
0,66	0,03505
0,67	0,03288
0,68	0,02545

2. *Clustering* menggunakan Metode *K-Means*

Jumlah *cluster* optimum adalah *cluster* memiliki nilai *Variance Ratio Criterion* paling tinggi. *Index* yang digunakan untuk penentuan jumlah *cluster* optimum dilakukan dengan *K* mulai dari 2 hingga 20.

Tabel 5.
Nilai VRC pada Layanan Ekspedisi JNE

Cluster	Nilai VRC	*Cluster*	Nilai VRC
2	3.239,8522	11	1.122,9674
3	2.304,4487	12	1.094,1687
4	1.949,2326	13	1.082,5548
5	1.751,9372	14	1.048,0313
6	1.613,7094	15	986,1774
7	1.459,8178	16	970,2019
8	1.391,1281	17	922,2360
9	1.278,7757	18	915,0433
10	1.182,0085	19	877,9982
20	854,2907		

Berdasarkan hasil nilai VRC pada Tabel 5 menunjukkan jumlah *cluster* (*K*) terbentuk sebanyak 20 dan yang memiliki nilai VRC paling tinggi yaitu sebesar 1.094,1687 pada 2 *cluster* dengan nilai *silhouette coefficient* sebesar 0,0819.

C. *Layanan Ekspedisi J&T*

Berikut adalah struktur data yang diperoleh setelah dilaku-kan *text prepocessing* pada akun layanan ekspedisi J&T yang ditampilkan pada Tabel 6.

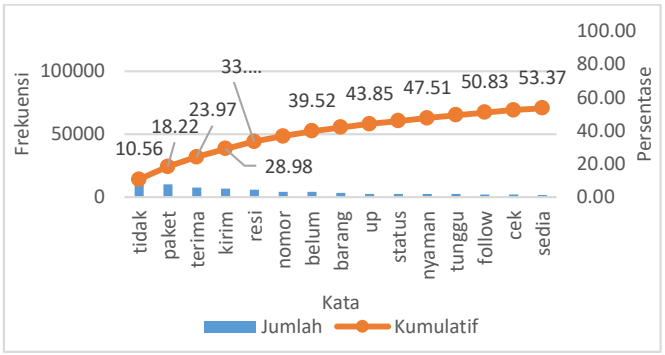
Tabel 6.
Struktur Data Layanan Ekspedisi J&T Setelah Dilakukan *Preprocessing*

Tweet ke-	Kata						
	barang	…	nama	…	status	…	terima
1	0	…	0	…	0	…	0
2	0	…	0	…	1	…	0
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
10.006	0	…	0	…	0	…	0
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
23.775	1	…	0	…	0	…	2

Jumlah dari frekuensi kemunculan setiap kata pada setiap *tweet* setelah dilakukan *preprocessing* yang ditampilkan pada Tabel 6 diketahui sebanyak 464 kata. Kata-kata tersebut yang akan dipergunakan sebagai variabel penelitian.

Gambar 7 adalah diagram pareto yang menunjukkan 30% kata yang sering muncul atau memiliki frekuensi ke-munculan yang paling besar dari total frekuensi kemunculan semua kata adalah kata "tidak", "paket", "terima", dan "kirim". Hal ini sesuai dengan hasil visualisasi *word cloud unigram* pada Gambar 6. Kata yang perlu diutamakan untuk ditindak lanjuti yaitu 30% kata yang muncul dari total frekuensi kemunculan kata. Kata "paket" menunjukkan barang atau produk dalam bungkusan yang dikirim oleh perusahaan ekspedisi. Kata "tidak" merupakan bentuk kata penolakan atau penyangkalan. Karena kata "paket" dan "tidak" tidak cukup memberikan informasi yang lebih dan lengkap, sehingga perlu dilakukan visualisasi *word cloud* menggunakan teknik *bigram*.

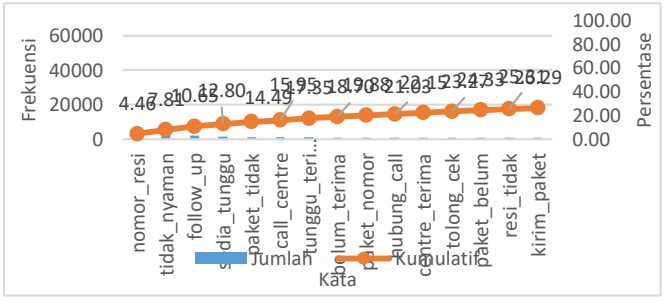
Gambar 6. Visualisasi *Word Cloud* pada Layanan Ekspedisi J&T.



Gambar 7 Diagram Pareto *Unigram* Layanan Ekspedisi J&T.

Berikut adalah visualisasi *word cloud* menggunakan teknik *bigram* pada akun layanan ekspedisi J&T yang ditampilkan pada Gambar 8.

Gambar 8. Visualisasi *Word Cloud Bigram* pada Akun Layanan Ekspedisi J&T.



Gambar 9. Diagram Pareto *Bigram* Layanan Ekspedisi J&T.

Berdasarkan diagram Pareto pada Gambar 9 diketahui bahwa 15% pasangan kata dari total frekuensi kemunculan kata adalah pasangan kata "nomor_resi", "tidak_nyaman", "*follow*_up", "sedia_tunggu", dan "paket_tidak". Pasangan kata tersebut sesuai dengan hasil visualisasi *word cloud bigram* pada Gambar 8. Kata yang perlu diutamakan untuk ditindak lanjuti adalah 15% pasangan kata dari total frekuensi kemunculan kata.

1. *Clustering* menggunakan Metode DBSCAN

Hasil *clustering* dengan metode DBSCAN dengan menggunakan *MinPts* sebesar 30 dengan berbagai kombinasi *Eps* memperoleh nilai *silhouette coefficient* untuk layanan ekspedisi JNE.

Tabel 8.
Nilai *Silhouette Coefficient* dari Metode DBSCAN

Eps	*Silhouette Coefficient*
0.1	0,9999464
0.2	0,9997138
0.3	0,9473227
0.4	0,7583361
0.5	0,3388869
0.6	0,1443148
0.7	0,0171568
0.8	0,0008924

Tabel 8 menunjukkan bahwa nilai *silhouette coefficient* tertinggi yang diperoleh dari metode DBSCAN pada akun layanan ekspedisi J&T dilakukan menggunakan parameter *Eps* sebesar 0,1 dilihat berdasarkan Tabel 8 adalah 0,9999. Nilai *silhouette coefficient* tersebut dapat dikatakan bahwa hasil *clustering* telah sangat baik karena *tweet* di dalam *cluster* yang ada telah kompak dan *tweet* dalam suatu *cluster* telah jauh dari *cluster* yang lain. Oleh karena itu, jumlah *cluster* optimum yang diperoleh dari *clustering* menggunakan metode DBSCAN dengan *MinPts* sebesar 30 dan *Eps* sebesar 0,1 adalah sebanyak 22 *cluster* dengan 20.384 *tweet* sebagai *noise*.

3. *Clustering* menggunakan Metode *K-Means*

Jumlah *cluster* optimum adalah *cluster* memiliki nilai *Variance Ratio Criterion* paling tinggi. *Index* yang digunakan untuk penentuan jumlah *cluster* optimum dilakukan dengan *K* mulai dari 2 hingga 20.

Tabel 9.
Nilai VRC pada Layanan Ekspedisi J&T

Cluster	Nilai VRC	*Cluster*	Nilai VRC
2	891,3695	12	424,9503
3	812,7370	13	423,6279
4	700,6291	14	410,3131
5	648,5735	15	382,1339
6	571,9525	16	381,3517
7	543,5358	17	364,3592
8	477,6850	18	347,6021
9	490,4468	19	360,1340
10	453,4703	20	345,5969
11	441,6000		

Berdasarkan hasil nilai VRC pada Tabel 9 menunjukkan jumlah *cluster* (*K*) terbentuk sebanyak 20 dan yang memiliki nilai VRC paling tinggi yaitu sebesar 891,3695 pada 2 *cluster* dengan nilai *silhouette coefficient* sebesar 0,04923.

D. Layanan Ekspedisi Pos Indonesia

Berikut adalah struktur data yang diperoleh setelah dilakukan *text prepocessing* pada akun layanan ekspedisi Pos Indonesia yang ditampilkan pada Tabel 10.

Tabel 10.
Struktur Data Layanan Ekspedisi Pos Indonesia Setelah Dilakukan *Preprocessing*

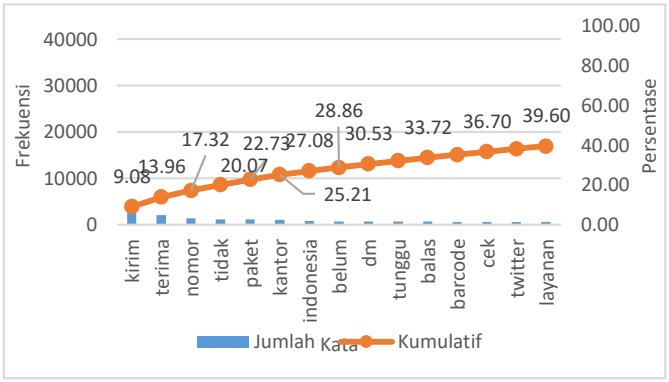
Tweet ke-	Kata						
	cek	…	kirim	…	paket	…	twitter
1	1	…	0	…	0	…	0
2	0	…	1	…	0	…	0
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮

Tabel 10.
Struktur Data Layanan Ekspedisi Pos Indonesia Setelah Dilakukan *Preprocessing* (Lanjutan)

Tweet ke-	Kata						
	cek	…	kirim	…	paket	…	twitter
3.000	0	…	3	…	2	…	0
⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
6.796	0	…	0	…	0	…	1

Jumlah dari frekuensi kemunculan setiap kata pada setiap *tweet* setelah dilakukan *preprocessing* yang ditampilkan pada Tabel 10 diketahui sebanyak 485 kata. Kata-kata tersebut yang akan dipergunakan sebagai variabel penelitian.

Gambar 10. Visualisasi *Word Cloud* pada Layanan Ekspedisi Pos Indonesia.

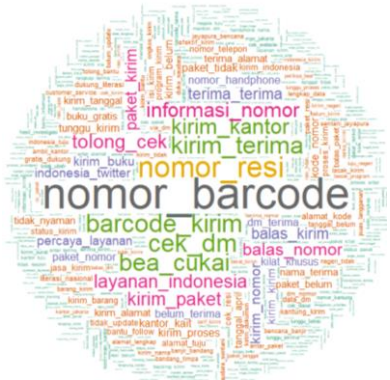


Gambar 11. Diagram Pareto *Unigram* Layanan Ekspedisi Pos Indonesia.

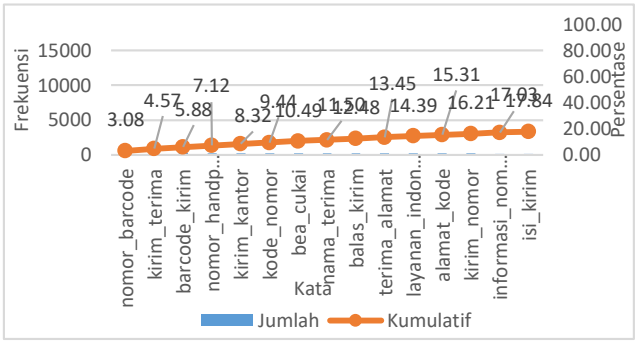
Berdasarkan diagram pareto pada Gambar 11 diketahui bahwa 30% kata yang sering muncul atau memiliki frekuensi kemunculan yang paling besar dari total frekuensi kemunculan semua kata adalah kata "kirim", "terima", "nomor", "tidak", dan "paket". Hal ini sesuai dengan hasil visualisasi *word cloud unigram* layanan Pos Indonesia pada Gambar 10. Kata perlu diutamakan untuk ditindak lanjuti yaitu 30% kata yang muncul dari total frekuensi kemunculan kata. Kata "kirim" merupakan kata yang memiliki frekuensi bobot yang paling besar. Kata "kirim" biasa digunakan untuk menanyakan lokasi paket. Kata tersebut tidak memiliki informasi yang lengkap, sehingga perlu informasi tambahan untuk dapat menginterpretasikan kata yang dihasilkan dari *word cloud*. Hasil visualisasi *word cloud* menggunakan teknik *bigram* yang ditampilkan pada Gambar 12.

Dari hasil diagram pareto pada Gambar 13 menunjuk-kan bahwa 10% pasangan kata dari total frekuensi ke-munculan kata adalah pasangan kata "nomor_barcode",

"kirim_terima", "barcode_kirim", "nomor_handphone", "kirim_kantor", dan "kode_nomor". Pasangan kata tersebut sesuai dengan hasil visualisasi *word cloud bigram* pada Gambar 12. Kata perlu diutamakan untuk ditindak lanjuti yaitu 10% kata yang muncul dari total frekuensi kemunculan kata.



Gambar 12. Visualisasi *Word Cloud Bigram* pada Akun Layanan Ekspedisi Pos Indonesia



Gambar 13. Diagram Pareto *Bigram* Layanan Ekspedisi Pos Indonesia.

1. *Clustering* menggunakan Metode DBSCAN

Hasil *clustering* dengan metode DBSCAN dengan menggunakan *MinPts* sebesar 30 dengan berbagai kombinasi *Eps* didapatkan nilai *silhouette coefficient* untuk layanan ekspedisi Pos Indonesia.

Tabel 12.
Nilai *Silhouette Coefficient* dari Metode DBSCAN

Eps	*Silhouette Coefficient*
0,1	0,9974903
0.2	0,9715307
0.3	0,9406119
0.4	0,8782036
0.5	0,7608612
0.6	0,3346245
0.7	0,2394187
0.8	0,0807623

Berdasarkan Tabel 12, nilai *silhouette coefficient* tertinggi yang diperoleh dari metode DBSCAN pada akun layanan ekspedisi Pos Indonesia dilakukan menggunakan parameter *Eps* sebesar 0,1. Nilai *silhouette coefficient* yang dihasilkan dengan menggunakan parameter *Eps* adalah 0,9974903. Hasil *clustering* dikatakan cukup baik dikarenakan nilai *silhouette coefficient* mendekati 1. Jumlah *cluster* yang diperoleh dari *clustering* menggunakan metode DBSCAN dengan *MinPts* sebesar 30 dan *Eps* sebesar 0,1 adalah sebanyak 11 *cluster* dengan 5.815 *tweet* sebagai *noise*.

2. *Clustering* menggunakan Metode *K-Means*

Jumlah *cluster* optimum adalah *cluster* memiliki nilai *Variance Ratio Criterion* paling tinggi. *Index* yang digunakan untuk penentuan jumlah *cluster* optimum dilakukan dengan *K* mulai dari 2 hingga 20.

Tabel 13.
Nilai VRC pada Layanan Ekspedisi Pos Indonesia

Cluster	VRC	*Cluster*	VRC
2	195,03808	12	117,2328
3	162,77934	13	116,1481
4	151,56805	14	111,8925
5	142,98908	15	108,8675
6	146,72316	16	110,7515
7	128,89546	17	104,8108
8	140,48242	18	103,5712
9	133,59514	19	102,1584
10	134,69985	20	94,41221
11	119,89227		

Berdasarkan hasil nilai VRC pada Tabel 13 menunjukkan jumlah *cluster* (*K*) terbentuk sebanyak 20 dan yang memiliki nilai VRC paling tinggi yaitu sebesar 195,03808 pada 2 *cluster* dengan nilai *silhouette coefficient* sebesar 0,0107.

E. Perbandingan Metode Clustering *DBSCAN dan* K-*Means*

Perbandingan metode *clustering* diperlukan untuk mengetahui metode yang terbaik yang dapat digunakan untuk mengelompokkan *tweet* pada akun layanan ekspedisi JNE, J&T, dan Pos Indonesia. Perbandingan metode *clustering* ini didasarkan pada hasil nilai *silhouette coefficient* yang didapatkan dari metode *K-Means* dan DBSCAN.

Tabel 14.
Perbandingan Antar Metode DBSCAN dan *K-Means*

Layanan Ekspedisi	Metode *Clustering*	*Silhouette Coefficient*	Jumlah *Cluster*
JNE	*K-Means*	0,0819	2
	DBSCAN	0,2652	18
J&T	*K-Means*	0,0492	2
	DBSCAN	0,9999	22
Pos Indonesia	*K-Means*	0,0107	2
	DBSCAN	0,9975	11

Berdasarkan Tabel 14 dapat dilihat dari nilai *silhouette coefficient* pada metode DSBCAN di semua layanan ekspedisi JNE, J&T, dan Pos Indonesia jauh lebih baik dalam me-ngelompokkan *tweet* dibandingkan dengan metode *K-Means*. Jumlah kelompok yang terbentuk banyak pada hasil *clustering* menggunakan metode DBSCAN dapat menguntung-kan pihak layanan ekspedisi JNE, J&T, dan Pos Indonesia untuk memberikan informasi yang lebih variatif agar dapat digunakan sebagai bahan evaluasi atau monitoring selanjutnya.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan analisis dan pembahasan, kata yang sering muncul pada *tweet* yang ditujukan pada layanan ekspedisi JNE adalah "cek" dan "dm". Kata yang sering muncul pada *tweet* yang ditujukan pada layanan ekspedisi J&T adalah "paket", sedangkan kata yang sering muncul pada *tweet* yang ditujukan pada layanan ekspedisi Pos Indonesia adalah "kirim". Berdasarkan nilai *silhouette coefficient* tertinggi yang diperoleh, didapatkan hasil bahwa metode DBSCAN merupakan metode terbaik dibandingkan dengan metode *K-Means* untuk mengelompokkan *tweet* yang ditujukan kepada akun media sosial *Twitter* layanan ekspedisi JNE, J&T, dan Pos Indonesia. *Clustering* dengan metode terbaik menghasilkan 18 cluster untuk layanan ekspedisi JNE, 22 *cluster* untuk layanan ekspedisi J&T, dan 11 *cluster* untuk layanan ekspedisi Pos Indonesia.

B. Saran

Saran yang dapat diberikan kepada layanan ekspedisi JNE, J&T, dan Pos Indonesia yaitu dapat mempertimbangkan hasil *clustering* dalam memberikan respon yang cepat kepada *tweet* pelanggan serta dapat dijadikan sebagai bahan evaluasi dan monitoring. Saran untuk penelitian selanjutnya, diharapkan agar lebih teliti pada saat *preprocessing* dan dapat meng-gunakan *feature selection genetic algorithm*.

DAFTAR PUSTAKA

[1] H. Tan, D. Plowman, and P. Hancock, "Intellectual Capital and Financial Returns of Companies," *J. Intellect. Cap.*, vol. 8, no. 1, 2007.

[2] S. Adinugroho and Y. Sari, *Implementasi Data Mining Menggunakan Weka*. Malang: UB Press, 2018.

[3] E. Irwansyah and M. Faisal, *Advanced Clustering : Teori dan Aplikasi*. Yogyakarta: Deepublish, 2015.

[4] R. Feldman and J. Sanger, *The Text Mining Handbook: Advance Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.

[5] Alfarisi, "Data Preprocessing-Konsep Pembelajaran Data Mining." steemit.com, 2017.

[6] G. Buntoro, T. Adji, and A. Purnamasari, "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation," in *Information Technology and Electrical Engineering (CITEE)*, 2014, pp. 39–43.

[7] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, *Text Mining Predictive Methods for Analyzing Unstructures Information*. New York: Spinger Science Business Media, Inc, 2005.

[8] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Bayesia Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. Sains dan Seni ITS*, vol. 4, no. 2, 2015.

[9] E. Dragut, F. Fang, P. Sistla, and C. Yu, *Stop Word and Related Problems in Web Interface Integration*. Chicago: University of Illinois, 2009.

[10] L. Bing, *Handbook of Natural Language Processing*. Boca Raton: CRC Press, 2010.

[11] T. Jo, *Text Mining : Concepts, Implementation, and Big Data Chalenge*. Seoul: Springer Internasional Publishing, 2018.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. Morgan Kaufman Publisher, 2012.

[13] Q. Ye, W. Gao, and W. Zeng, "Color Image Segmentation Using Density-Based Clustering," in *International Conference on Multimedia and Expo (ICME)*, 2003, p. 346.

[14] S. Thomas and U. Harode, "A Comparative Study on K-Means and Hierarchical Clustering," *Int. J. Electron. Electr. Comput. Syst.*, vol. 4, pp. 5–11, 2015.

[15] T. Calinski and J. Harabasz, "A Dendritc Method for Cluster Analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.

[16] J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*. New York: Cambrige University Press, 2007.

[17] T. Kwartler, *Text Mining in Practice with R*. USA: John Wiley & Sons, 2017.

[18] Rapi, "Mengenal Sedikit Mengenai Perusahaan Jasa Pengiriman Barang," *rapi.co.id*, 2017.