Implementasi *Text Mining* Untuk Pengelompokan Ulasan Pelanggan *E-Commerce* Berdasarkan Topik Ulasan Menggunakan Algoritma *Hierarchical Agglomerative Clustering*

Li'Izza Diana Manzil, Prof. Dr Mohammad Isa Irawan, M.T.
Matematika, *Fakultas Matematika, Komputasi dan Sains Data, Institut Teknologi Sepuluh Nopember (ITS)*
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: mii@its.ac.id

***Abstrak*— Pesatnya perkembangan teknologi dalam bidang bisnis memunculkan banyak *e-commerce* di Indonesia. Kini sebagian besar masyarakat lebih suka melakukan pembelian melalui internet, karena lebih mudah dalam transaksinya. Namun, pembelian barang melalui *e-commerce* terkadang tidak sesuai dengan apa yang diinginkan, baik secara pelayanan maupun barang yang diperoleh. Oleh karena itu, banyak *e-commerce* yang memberikan fasilitas pemberian ulasan mengenai pelayanan ataupun barang. Ulasan yang diberikan pelanggan sangat beragam. Dengan demikian, perlu adanya suatu penelitian untuk mengelompokkan dan mengklasifikasikan ulasan untuk memudahkan penjual dan calon pelanggan *e-commerce* untuk mengetahui aspek apa yang diulas oleh pelanggan. Metode yang digunakan pada penelitian ini adalah clustering dengan algoritma *Hierarchical Agglomerative Clustering (HAC)*. Pada penelitian ini, dilakukan pemotongan jarak hasil *clustering* pada *threshold* 0.94, 0.945 dan 0.95. Hasil perhitungan precision dari clustering dengan beberapa *threshold* tersebut menyatakan bahwa clustering dengan *threshold* 0.96 lebih baik daripada menggunakan *threshold* lainnya dengan nilai presisi 85.2%**
Kata Kunci: *E-commerce, HAC, Ulasan*

I. PENDAHULUAN

ERA yang serba digital membuat teknologi memberikan dampak pada gaya hidup masyarakat. Pesatnya perkembangan teknologi banyak memberikan kemudahan bagi masyarakat. Salah satu contoh teknologi adalah internet. Internet telah menjadi kebutuhan primer bagi masyarakat. Berbagai aspek kehidupan telah dipengaruhi oleh internet. Seperti dalam bidang bisnis, banyak bermunculan bisnis online dan maraknya kegiatan berbelanja melalui media internet. Suatu survei yang telah dilakukan oleh APJII pada tahun 2016 menyebutkan bahwa terdapat 98,6% pengguna internet di Indonesia mengetahui internet sebagai tempat jual beli barang dan jasa. Sementara 63,5% atau sekitar 84,2 juta pengguna internet pernah melakukan transaksi online. Hal itu ditandai dengan banyaknya *e-commerce* yang muncul dan semakin berkembang.

Hadirnya *e-commerce* membuat masyarakat sering berbelanja di toko online, daripada dating ke toko offline langsung. Hanya dengan melakukan kegiatan sederhana seseorang sudah bisa membeli barang yang diinginkan. Namun hal itu juga mempunyai kekurangan. Orang yang akan membeli barang secara online tidak bisa mengetahui bagaimana kualitas barang yang diinginkan, sehingga tidak jarang juga pembeli yang merasa kecewa dengan barang yang telah dibeli. Oleh karena itu, beberapa *e-commerce* menyediakan fasilitas untuk memberikan komentar mengenai barang yang dibeli atau pelayanan yang telah diberikan. Komentar tersebut dinamakan ulasan.

Ulasan yang diberikan pelanggan sangat beragam. Ulasan yang diberikan bisa berupa pujian atau kritikan. Banyaknya ulasan yang diberikan pelanggan seringkali membuat penjual merasa kesulitan dalam menyimpulkan aspek yang dibahas pada ulasan. Untuk mengolah dan memantau ulasan yang diberikan pelanggan bukan hal yang mudah karena jumlah ulasan yang dimuat dalam website umumnya sangat banyak apabila dilihat secara manual.

Dalam penelitian ini akan dilakukan proses text mining untuk pengelompokan ulasan pelanggan berdasarkan aspek yang dibahas. Pengelompokan ulasan menggunakan salah satu metode pada clustering dan klasifikasi. Metode *clustering* yang akan digunakan algoritma *Hirarchical Agglomerative Clustering* (HAC). Proses *clustering* pada penelitian ini bertujuan untuk mengelompokkan ulasan dan menentukan topik yang sedang dibahas pada ulasan. Sedangkan, proses klasifikasi digunakan untuk mengelompokkan ulasan-ulasan pelanggan berdasarkan topik yang telah dihasilkan dari proses *clustering.*

II. TINJAUAN PUSTAKA

A. *E-Commrce*

E-commerce diartikan sebagai penggunaan internet dan *website* untuk bertransaksi bisnis atau bisa juga didefinisikan sebagai suatu transaksi perdagangan secara digital antar organisasi atau individu [1]. *E-commerce* merupakan transaksi digital antara dan diantara organisasi atau secara individu [2]. Perusahaan *e-commerce* juga melakukan kegiatan yang mendukung transaksi pasar mereka, seperti periklanan, pemasaran, dukungan pelanggan, kemanan, pengiriman dan pembayaran [2].

B. *Text Mining*

Text mining merupakan suatu proses menemukan dan mengekstraksi informasi dari sekumpulan sumber text yang banyak dan tidak terstruktur [3]. Text mining yaitu suatu

proses ekstraksi pengetahuan implisit dari data tekstual [4]. Adapun langkah-langkah yang dilakukan dalam *text mining* adalah:

1. Praproses Data

Pada tahap ini terdapat beberapa proses, seperti penghilangan tanda baca, tokenisasi, *stopword removal*, dan *stemming*. Tokenisasi merupakan pemisahan kata berdasarkan spasi antar kata. Kemudian, *stopword removal* merupakan penghilangan kata yang kurang penting, misalnya kata ”yang”, ”dan”, ”tapi”, dan lain-lain. Keberadan kata-kata tersebut akan berpengaruh terhadap hasil yang diberikan. Selain itu, *stemming* adalah mengubah kata menjadi kata dasarnya.

2. Text Transformation

Pada tahap ini dilakukan proses transformasi dari teks ke angka. Proses transformasi ini bisa juga disebut dengan pembobotan kata. Pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Proses TF-IDF terbentuk dari dua proses yaitu TF dan IDF. Adapun rumus dalam mencari bobot TF adalah:

$$tf_{i,j} = t_{i,j} \quad (2.1)$$

dengan:

$tf_{i,j}$: *Term frequency* (jumlah kemunculan suatu kata i dalam suatu dokumen ke-j)

$t_{i,j}$: Jumlah kata i dalam dokumen j

sedangkan rumus IDF adalah sebagai berikut:

$$idf_i = log\left(\frac{N}{doc_i}\right) + 1 \quad (2.2)$$

dengan:

idf_i : inverse document frequency kata i

N : jumlah dokumen yang digunakan

doc_i : jumlah dokumen yang mengandung kata *i*

Dari rumus 2.1 dan rumus 2.2 didapatkan nilai bobot TF-IDF untuk kata i pada dokumen ke- j adalah sebagai berikut:

$$w_{i,j} = tf_{i,j} \times idf_i \quad (2.3)$$

Normalisasi bobot TF-IDF dari hasil pada rumus 2.3 dapat dilakukan menggunakan persamaan 2.4 berkut:

$$w_{i,j} = \frac{w_{i,j(pra)}}{\left|\sum_{j=1}^{n} w_{i,j(pra)}\right|} \quad (2.4)$$

3. Feature Selection

Feature selection adalah tahap lanjut dari pengurangan dimensi pada proses text transformation. Proses ini bertujuan untuk pemilihan kata-kata relavan yang benar-benar mempresentasikan isi dari suatu dokumen.

4. Pattern Discovery

Pattern Discovery merupakan tahap penting untuk menemukan pola dari keseluruhan teks. Dalam penemuan Dalam penemuan pola ini, proses text mining dikombinasikan dengan proses-proses data mining. Masukan awal dari proses text mining adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai evaluasi. Terdapat beberapa jenis *pattern discovery* diantaranya klasifikasi, *clustering*, asosiasi, dan analisa tren [5].

C. Latent Semantic Analysis

Latent Semantic Analysis (LSA) merupakan metode untuk menemukan korelasi, hubungan, kemiripin, atau keterkaitan antar dokumen, penggalan dokumen, dan kata-kata yang muncul pada dokumen dengan pendekatan matematika [6]. LSA yaitu suatu teknik matematika atau statistika untuk mengekstraksi dokumen dalam menemukan hubungan dari penggunaan kata secara kontekstual. Hal itu bukan termasuk pemrosesan bahasa alami yang sederhana maupun kecerdasan buatan.

D. Singular Value Decomposition

Singular Value Decomposition (SVD) merupakan suatu metode untuk mengidentifikasi dan mengurutkan dimensi yang menunjukkan data mana yang mempunyai variasi paling banyak. Berkaitan dengan hal tersebut, SVD dapat mengidentifikasi dimana variasi yang muncul paling banyak, sehingga hal ini memungkinkan untuk mencari pendekatan yang terbaik pada data asli menggunakan dimensi yang lebih kecil. Adapun matriks SVD ditunjukkan pada persamaan (2.5):

$$A = U\Sigma V^T \quad (2.5)$$

E. Cosine Similarity

Metode *Cosine Similarity* merupakan metode untuk mencari jarak antar dokumen untuk mengetahui tingkat kesamaannya. Secara umum perhitungan metode ini didasarkan pada *vector space similarity measure.* Metode *cosine similarity* ini menghitung similarity antara dua buah ulasan yang dinyatakan dalam dua vektor. Untuk memperoleh *cosine similarity* dapat menggunakan persamaan berikut

$$Cos(d_1, d_2) = \frac{d_1 . d_2}{\|d_1\| \|d_2\|} \quad (2.6)$$

dengan:

$\cos(d_1, d_2)$: nilai *cosine similarity* antara dokumen 1 dan dokumen 2

Nilai yang dihasilkan dari perhitungan menggunakan persamaan 2.6 menhasilkan nilai pada interval -1 sampai 1. Hal ini bertentangan dengan bag of word yang nilainya berada pada rentang 0 hingga 1 karena vektor frekuensi tidak pernah negative. Sehingga Jarak antara dua dokumen dapat menggunakan persamaan berikut.

$$dis(d_1,d_2) = 1 - \cos(d_1,d_2) \quad (2.7)$$

dengan:
$dis(d_1,d_2)$: jarak antara dokumen 1 dan dokumen 2

F. *Hierarchical Agglomerative Clustering*

Hierarchical clustering merupakan salah satu algoritma pada proses clustering yang mengelompok berdasarkan tingkatan tertentu. Hierarchical algoritma diantara *top-down* atau bottom-up. Algoritma *bottom-up* memperlakukan setiap dokumen yang awalnya ter-cluster tunggal yang kemudian secara berturut-turut menggabungkan (atau mengelompokkan) pasangan cluster sampai semua cluster telah bergabung menjadi satu cluster yang berisi semua dokumen. Algoritma *bottom-up* ini kemudian disebut dengan *Hierarchical Agglomerative Clustering* (HAC).

Berikut merupakan algoritma *Hierarchical Agglomerative Clustering* untuk mengelompokkan N objek [1]:

1. Mulai dengan *N* cluster, setiap cluster mengandung entity tunggal dan sebuah matriks simetri dari jarak $D = \{d_{ik}\}$ dengan tipe matrik adalah NxN.
2. Mencari matriks jarak untuk pasangan cluster yang terdekat, yaitu dengan mencari jarak terbesar. Misalkan jarak antara cluster d_{ab}.
3. Gabungkan cluster A dan B menjadi label AB yang baru. Update inputan pada matrik jarak dengan cara:

a. Hapus baris dan kolom yang bersesuaian cluster A dan B
b. Tambahkan baris dan kolom yang memberikan jarak-jarak antara cluster (AB) dan cluster-cluster yang tersisa.

Beberapa metode hierarchical clustering yang sering digunakan yaitu *single linkage, complete linkage, average linkage, ward's linkage, centroid linkage*.

G. *Average Linkage*

Average linkage merupakan proses pengelompokan dokumen berdasarkan pada jarak rata-rata antar objeknya. Untuk menghitung jarak antar dokumen menggunakan perssmaan (2.7) berikut:

$$dist_{avg}(C_i,C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i p^0 \in C_j} |p - p^0| \quad (2.7)$$

dengan:
$dist_{avg}(C_i,C_j)$: rata-rata jarak antara cluster i dengan cluster j
n_i : banyaknya objek dalam cluster i
n_j : banyaknya objek dalam cluster j
$|p - p^0|$: jarak antara objek p dan p^0

III. METODOLOGI PENELITIAN

A. *Metode Penelitian*

Pada penelitian Tugas Akhir ini akan dilakukan penelitian berdasarkan langkah-langkah sebagai berikut:

a. Pengumpulan data

Pada tahap ini dilakukan penelitian data-data ulasan pelanggan Tokopedia terhadap suatu produk. Data ulasan tersebut diperoleh dari hasil *scraping* dari halaman situs *website* www.tokopedia.com

b. Praproses data

Pada tahap praproses data ini dilakukan dalam tiga tahap yaitu tokenisasi, *stopword removal* dan *stemming*.

c. Pembuatan TF-IDF

Pembuatan TF-IDF memberikan bobot terhadap data yang dihasilkan diproses sebelumnya dengan menggunakan persamaan 2.1. Langkah pertama yang harus dilakukan yaitu membuat skema TF-IDF. Dari proses tersebut akan dihasilkan suatu matriks dokumen-kata (*document-term matrix*).

d. Identifikasi kemiripan dokumen

Identifikasi kemiripan dokumen merupakan salah satu kegiatan *feature selection*. Identifikasi kemiripan dokumen ini bertujuan untuk mengetahui adanya kemiripan atau hubungan antar dokumen. Tahap ini menggunakan metode LSA untuk memodelkan matriks dengan memanfaatkan matriks dari langkah (c). Dalam implementasinya, matriks DTM akan direduksi menjadi matriks $U\Sigma V^T$, kemudian dilakukan *truncated SVD* untuk masing-masing k yang dikehendaki.

e. Proses *Clustering*

Tahap selanjutnya data akan di-*cluster* menggunakan algoritma *Hierarchical Agglomerative Clustering* dengan metode *average linkage* untuk mengelompokkan dokumen dan membentuk suatu topik. Proses clustering dimulai dengan menghitung jarak antar dokumen dengan menggunakan persamaan (2.6). Hasil perhitungan jarak kemudian digunakan oleh sistem untuk melakukan proses clustering dengan *hierarchical agglomerative clustering*.

f. Analisa Hasil Clustering

Pada tahap ini, akan dilakukan analisa mengenai hasil clustering dan klasifikasi untuk mengetahui performansi dari metode yang digunakan untuk pengelompokan topik ulasan pelanggan Untuk menuliskan label pada sumbu-sumbu dari sebuah diagram/gambar lebih baik digunakan kata daripada simbol. Pastikan semua simbol maupun kata dapat dibaca (*readable*).

g. Penarikan Kesimpulan

Pada tahap ini akan dilakukan penarikan kesimpulan dari penelitian yang telah dijalankan, yakni mengenai aspek apa saja yang diulas oleh pelanggan berdasarkan hasil analisa dan akurasi metode yang telah digunakan/

IV. IMPLEMENTASI DAN HASIL

A. *Deskripsi Data*

Pada uji coba ini data yang digunakan merupakan data ulasan suatu produk yang ada di www.tokopedia.com. Data diambil dengan cara *scraping* sebagaimana yang telah dijelaskan pada bab sebelumnya. Jumlah data yang diambil adalah 2449 ulasan dari produk tas. Ulasan-ulasan tersebut disimpan dalam bentuk *.csv.*

B. *Praproses Data*

Praproses dilakukan untuk memperoleh data bersih yang dapat digunakan untuk proses selanjutnya. Praproses yang dilakukan adalah *tokenization, stopword removal* dan *stemming.* Dari tahap-tahap tersebut data yang dihasilkan menjadi 1793 ulasan.

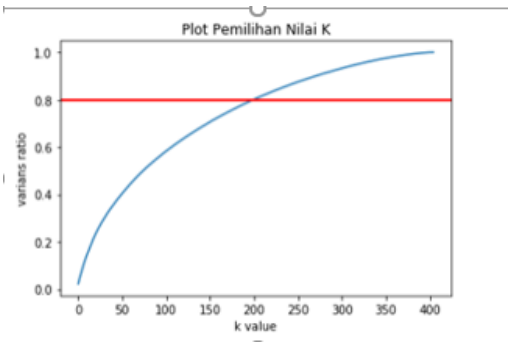
C. *Perhitungan TF-IDF*

Setelah data telah bersih, dilakukan perhitungan tf-idf menggunakan persamaan 2.3 dan dinormalisasi dengan menggunakan rumus 2.4

Tahap ini diawali dengan pembuatan *document-by-term* untuk menunjukkan frekuensi tiap kata pada tiap ulasan, dimana baris matriks menunjukkan ulasan dan kolom menunjukkan kata-kata. Setelah itu dilakukan normalisasi agar proses komputasi berjalan lebih cepat.

D. Penentuan Jumlah *Principal Components*

Pada tahap selanjutnya dilakukan pemotongan matriks menjadi matriks *principal components*. Jumlah *principal components* ditentukan dengan menggunakan 0.8 sebagai titik pemilihan nilai *k* untuk *Truncated SVD.* Dengan melihat grafik pada Gambar 5.4 dapat diketahui interval jumlah *principal components.* Pada grafik tersebut dapat diketahui bahwa dengan menggunakan threshold 0.8 ditemukan interval nilai *principal components* 200 sampai 450, namun tidak diketahui angka pastinya. Oleh karena itu, dengan menggunakan *syntax* berikut akan ditemukan jumlah *principal components.*



Gambar 5. 1 Grafik Penentuan Jumlah *Principal Components*

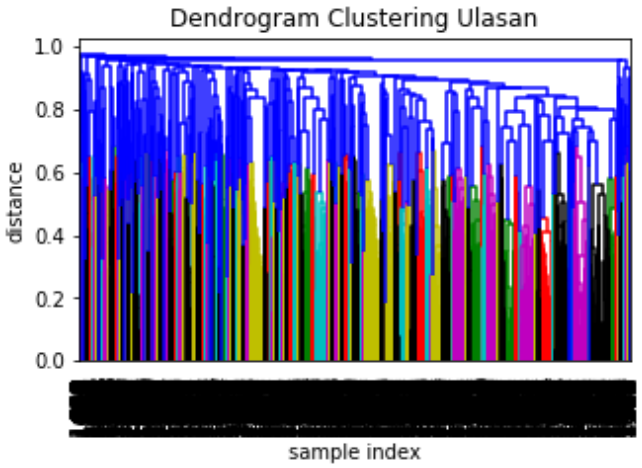
Nilai *principal components* dapat dilihat pada kutipan matriks Gambar 5.2 berikut.

	0	1	2	3	4	5	6	7	8
0	0.000000	0.999823	0.999516	0.789015	0.999963	1.000394	0.999854	0.999953	0.999515
1	0.999823	0.000000	0.921049	0.836551	1.000132	1.000443	1.000611	0.835192	0.823845
2	0.999516	0.921049	0.000000	0.964478	1.000072	1.000763	0.996010	1.000016	0.937699
3	0.789015	0.836551	0.964478	0.000000	0.950315	0.997968	1.001766	0.694355	0.862664
4	0.999963	1.000132	1.000072	0.950315	0.000000	0.999994	0.999964	0.837009	0.924298
5	1.000394	1.000443	1.000763	0.997968	0.999994	0.000000	0.999296	0.999900	1.002711
6	0.999854	1.000611	0.996010	1.001766	0.999964	0.999296	0.000000	1.000030	0.662411
7	0.999953	0.835192	1.000016	0.694355	0.837009	0.999900	1.000030	0.000000	0.899971
8	0.999515	0.823845	0.937699	0.862664	0.924298	1.002711	0.662411	0.899971	0.000000
9	0.999954	1.000039	0.999833	0.999345	1.000104	1.000054	0.999779	0.999987	1.001852
10	0.999923	0.						411	0.825133
11	0.759284	1.						098	1.000619
12	1.000014	0.834718	0.999794	0.887071	0.943524	1.000349	1.000636	0.722735	0.923579

Gambar 5. 2 Kutipna Principal Components Matrix

5.2.4 *Pengelompokan Menggunakan HAC*

Pada proses ini akan dilakukan proses clustering ulasan. Clustering yang dihasilkan dengan menggunakan metode HAC membentuk satu cluster besar. Berikut ini akan ditunjukkan hasil *clustering* pada gambar 5.3.



Gambar 5. 3 Dendrogram Hasil Clustering

Untuk menentukan kelompok-kelompok ulasan yang memiliki topik yang sama dibutuhkan titik *threshold* agar terbentuk menjadi beberapa cluster. Pada pengujian ini akan dilakukan clustering ulasan dengan menggunakan beberapa *threshold*. Dari tiap nilai *threshold* yang digunakan memiliki jumlah cluster yang berbeda-beda. Nilai *threshold* yang digunakan adalah 0.94, 0.95 dan 0.96. Berikut penjelasan tiap *clustering.*

a. Clustering dengan *Threshold* 0.94

Dengan *threshold* 0.94 proses *clustering* menghasilkan 15 topik *cluster* dengan banyak ulasan dan tiga *keyword* yang dihasilkan dapat dilihat pada tabel 5.1 berikut.

Tabel 5.1 Hasil Clustering dengan *Threshold* 0.94

Cluster	Jumlah Ulasan	*Keyword*
1	2	‘jalan’,’jalan jalan’, ‘lumayan’
2	8	‘selamat’, ’darat’, ‘bagus’
3	43	‘deskripsi’,’gambar’,’gambar deskripsi’
4	3	‘bagus’,’lanjut’,’cuma’
5	4	‘bintang’, ‘bintang bicara’,’bicara’
6	53	‘ragu’,’jual’,’sama’
7	5	‘bagus’, ‘terjangkau bagus’, ‘harga terjangkau’
8	48	‘jelek’, ‘hati’,’kualitas bahan’
9	11	‘cacat’,’mantap’,’sesal’
10	6	‘cocok’,’pas’,’mantap’
11	17	‘aman’,’cepat’,’cepat aman’
12	1593	‘sesuai’,’warna’,’pesan’
13	1434	‘sesuai’,’cepat’,’bagus’
14	24	‘resleting’,’resleting rusak’,’rusak’
15	55	‘lambat’,’kirim’,’lama’

Berdasarkan tabel 5.1 diatas mayoritas ulasan tergabung dalam *cluster* 13 dengan banyak ulasan adalah 1434. Sedangkan *cluster* dengan anggota paling sedikit adalah pada *cluster* 1 dan 3 yaitu sebanyak 2 ulasan.

Topik yang dibahas pada tiap ulasan dapat ditentukan berdasarkan keyword yang dihasilkan. Untuk topik tiap kelompok ulasan hasil *clustering* dapat dilihat pada tabel berikut.

Tabel 5. 1 Topik Ulasan Hasil Clustering Threshold 0.94

Cluster	Topik
1	Lumayan untuk jalan-jalan
2	Barang sampai
3	Deskripsi gambar
4	Lain-lain
5	Rating
6	Keraguan sama penjual
7	Harga Produk
8	Kualitas bahan produk
9	Kekecewaan Pelanggan
10	Kesesuaian
11	Efektifitas pengiriman
12	Kesesuaian warna
13	Kesesuaian dan kecepatan
14	Resleting produk
15	Pengiriman lambat

b. Clustering dengan *Threshold* 0.95

Pemotongan jarak *cluster* pada titik 0.95 menghasilkan 9 cluster. Pada *cluster* ini terdapat penggabungan *cluster* 7 dengan cluster 8 dan cluster 10 sampai *cluster* 15 pada hasil *cluster* dengan threshold 0.94 ke *cluster* 9. Sehingga hasil *cluster* dengan threshold 0.95 dapat dilihat pada Tabel 5.3 berikut.

Tabel 5.3 Hasil Clustering dengan *Threshold* 0.95

Cluster	Jumlah Ulasan	*Keyword*
1	2	‘jalan’,’jalan jalan’, ‘lumayan’
2	32	‘selamat’, ’darat’, ‘bagus’
3	2	‘deskripsi’,’gambar’,’gambar deskripsi’
4	20	‘bagus’,’lanjut’,’cuma’
5	53	‘bintang’,’bintang bicara’,’bicara’
6	7	‘ragu’,’jual’,’sama’
7	8	‘bagus’, ‘harga’,’kualitas’
8	18	‘cacat’,’mantap’,’sesal’
9	1651	‘sesuai’,’cepat’,’bagus’

Berdasarkan tabel 5.3 jumlah ulasan pada *cluster* 1 sampai *cluster* 6 sama dengan hasil *cluster* dengan *threshold* 0.94. Sedangkan untuk *cluster* dengan *keyword* ‘sesuai’, ‘cepat’, dan ‘bagus’ mengalami peningkatan

menjadi 1651 ulasan. Untuk topik setiap ulasan dapat dilihat pada Tabel 5.4 berikut.

Tabel 5. 4 Topik Ulasan Hasil Clustering Threshold 0.95

Cluster	Topik
1	Lumayan untuk jalan-jalan
2	Barang sampai
3	Deskripsi gambar
4	Lain-lain
5	Rating
6	Keraguan sama penjual
7	Kualitas dan harga
8	Kekecewaan pelanggan
9	Kesesuaian dan kecepatan

Pada *cluster* ini terdapat perbedaan topik dengan proses clusterimg sebelumnya. Hasil *cluster* ini terdapat ulasan yang membahas kualitas dan harga, sedangkan pada proses *cluster* sebelumnya kualitas dan harga terletak pada cluster yang berbeda. Hal itu terdapat penggabungan anggota cluster dengan cluster lainnya.

c. Clustering dengan *Threshold 0.96*

Proses *cluster* dengan *threshold* 0.96 diperoleh hasil *cluster* sebagai berikut Tabel 5.5 berikut.

Tabel 5.5 Hasil Clustering dengan *Threshold* 0.96

Cluster	Jumlah Ulasan	*Keyword*
1	2	‘jalan’,’jalan jalan’, ‘lumayan’
2	32	‘selamat’, ’darat’, ‘bagus’
3	2	‘deskripsi’,’gambar’,’gambar deskripsi’
4	1757	‘sesuai’,’cepat’,’bagus’

Berdasarkan tabel 5.5 diatas, dihasilkan 4 jumlah cluster dimana setiap clusternya memiliki 2, 32, 2 dan 1757 ulasan. Pada proses *cluster* dengan threshold 0.96 cluster ke 1, 2, dan 3 memiliki hasil cluster yang sama dengan hasil *cluster* dengan *threshold* 0.96. Topik-topik cluster dapat dilihat pada Tabel 5.6 di bawah ini.

Tabel 5. 6 Topik Ulasan Hasil Clustering Threshold 0.96

Cluster	Topik
1	Lumayan untuk jalan-jalan
2	Barang Sampai
3	Deskripsi gambar
4	Kesesuaian dan kecepatan

Dari ketiga hasil *cluster* dengan *threshold* - *threshold* tersebut beberapa keyword yang muncul beberapa kali di beberapa cluster. Hal itu dikarenakan kata tersebut memiliki frekuensi kemunculan kata yang tinggi yang juga dapat mempengaruhi hasil perhitungan tf-idf tiap cluster

E. Analisa Hasil Clustering

Analisis hasil *cluster* digunakan untuk mengetahui *threshold* mana yang lebih optimal diantara penggunaan *threshold* lainnya. Hal itu dilakukan dengan menghitung rata-rata *presisi* dari setiap proses *cluster*. Untuk perhitungan *presisi* pada tiap *cluster* dengan menggunakan rumus berikut.

$$presisi_i = \frac{jumlah\ ulasan\ yang\ sesuai}{jumlah\ ulasan\ hasil\ clustering}$$

Proses perhitungan presisi dilakukan dengan menghitung presisi tiap *cluster*. Kemudian untuk mengahsilkan presisi proses *cluster* dengan menghitung rata-rata presisi tiap cluster. Berikut merupakan hasil presisi dari setiap *cluster*.

Table 5.7 Nilai Presisi Tiap Threshold

No	*Threshold*	Presisi
1	0.94	80.2%
2	0.95	79.1%
3	0.96	85.2%

Beradasrkan Tabel 5.7 nilai presisi dari proses *cluster* dengan *threshold* 0.96 adalah 85.2%. Hal itu berarti 85.2% ulasan telah tergabung pada cluster yang sesuai.

Dari perhitungan ketiga proses *cluster* tersebut dapat diketahui bahwa *cluster* dengan *threshold* 0.96 lebih tinggi dibandingkan dengan proses cluster dengan *threshold* lainnya. Hal itu dikarenakan tiap cluster memiliki ulasan yang memiliki topik yang sesuai.

V. KESIMPULAN

Kesimpulan yang dapat ditarik dari penelitian ini adalah sebagai berikut:

1. Implementasi *text mining* untuk mengelompokan ulasan dimulai dengan melakukan praproses data. Setelah data bersih dilakukan perhitungan TF-IDF sebagai proses pembobotan kata. Hasil dari proses TF-IDF merupakan matriks *document-by-term.* Matriks tersebut kemudian diuraikan menggunakan SVD. Hasil penguraian tersebut dilakukan pereduksian dimensi matriks sebesar *k* sehingga menjadi *principal components matrix.* Kemudian melakukan *cluster* dengan algoritma *Hierarcical Agglomerative Cluster* pada *principal components matrix*. Sebelum dilakukan *cluster*, terlebih dahulu mencari jarak tiap dokumen dengan menggunakan *cosines similarity* untuk mengetahui kemiripan dokumen.
2. Berdasarkan perhitungan presisi, *cluster* dengan *threshold* 0.96 memiliki nilai presisi yang lebih tinggi daripada dengan menggunakan *threshold* yang lainnya. Nilai presisi proses *cluster* tersebut adalah 85.2%.

DAFTAR PUSTAKA

[1] Griva, A. 2018. “Retail Business Analytics : Computer Visit Segmentation using Market Basket Data,”

[2] Nugroho, A. 2006. E-Commerce - Memahami Perdagangan Modern Di Dunia Maya, Bandung: Informatika.

[3] KM, Shivaprasad, T. Hanumantha Reddy. 2016, “Text Mining : An Improvised Feature Based Model Approach,” in 2nd International Conference on Applied and Theoritical Computing and Communication Technology.

[4] Feldman, Ronen & James Sanger. 2007. "Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data". New York: Cambridge.

[5] Nugroho, E. 2011. "Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp", Malang: Universitas Brawijaya.

[6] Wicaksono, D.W. 2014. “Sistem Deteksi Kemiripan Dokumen Teks Menggunakan Model Bayesian Pada Term Latent Semantic Analysis,” Tugas Akhir, Institut Teknologi Sepuluh Nopember, Surabaya

[7] Baker, K., “*Singular Value Decomposition Tutorial*,” 2005. [Online]. Available: davetang,org.

[8] Berry, M.W, S.T. Dumais, G.W. O'Brien. 1994. “*Using Linear Algebra for Intelligent Information Retrieval,”*

[9] Han, Jiawei, Micheline Kamber, Jian Pei. 2012. "Data Mining : Concepts and Techniques 3rd edition". Waltham: Morgan Kauffman PublisherN. Kawasaki,