CLUSTERING PADA DATA SENTIMENT PENGGUNAAN TRANSPORTASI ONLINE MENGGUNAKAN ALGORITMA SINGLE PASS CLUSTERING

CLUSTERING ON SENTIMENT DATA ONLINE TRANSPORTATION USING SINGLE PASS CLUSTERING ALGORITHM

Ja'far Razzaq[1], Fairuz Azmi[2], Casi Setianingsih[3]
[1,2,3] Universitas Telkom, Bandung
jrazzaq@student.telkomuniversity.ac.id1, **worldliner@telkomuniversity.ac.id[2]**,
setiacasie@telkomuniversity.ac.id[3]

Abstrak

Transportasi *online* hadir ditengah – tengah masyarat sebagai solusi untuk masyarakat saat ingin berpergian, namun sangat padatnya pengguna transportasi umum. Namun, dengan semakin menjamurnya transportasi *online,* semakin banyak juga komentar yang diberikan oleh penggunanya mengenai kinerja dari transportasi *online* baik itu komentar posistif, negatif, atau netral. Pada penelitian ini penulis mencoba melakukan pengelompokan data terhadap komentar positif, negatif, dan netral yang dimana komentar diambil melalui media sosial Instagram penyedia jasa transportasi *online* Go-Jek Indonesia dan Grab Indonesia. Data yang telah didapat kemudian melalui tahapan *pre-processing*, pembobotan kata, *clustering* dimana pada penelitian ini menggunakan algoritma *Single Pass Clustering*, dan kemudian hasil *clustering* ditampilkan di *website*. Hasil pengujian dari proses *clustering* dengan mengguankan *threshold* 0.1 sampai 0.9 didapat bahwa semakin besar nilai *threshold* semkin cepat proses *clustering* dan semakin sedikit *clusternya*. Hasil terbaik didapat pada *threshold* 0.1 dimana pada dataset positif didapat hasil 123 *cluster* dengan kecepatan 0.001196800s. Sedangkan pada dataset negatif didapat hasil 170 *cluster* dengan kecepatan 0.002018130s. Sedangkan pada dataset netral didapat hasil 151 *cluster* dengan kecepatan 0.001530701s.

Kata kunci : Transportasi Online, Sosial Media Instagram, Clustering, Single Pass Clustering

Abstract

Online transportation is present in the midst of society as a solution for people when they want to trevel, but the number of users of public transportation is very dense. However, with the proliferation od online transportation, more and more comments are given by users regrading online transportation, whether they are positive, negative, or neutral comments. In this study, are taken through Instagram social media, online transportation service providers Go-Jek Indonesia and Grab Indonesia. The data that has been obtained then goes through the stages of pre-processing, word weighting, clustering which in this study uses the Single Pass Clustering algorithm, and then the clustering result are displated on the website. The test result from the clustering process using a threshold of 0.1 to 0.9, it is found that the greater threshold value, the faster clustering process and the fewer clusters. With the best cluster results found at the threshold 0.1 in the positive datasets, the results obtained are 123 clusters with a clustering speed of 0.001196800s, while in the negative datasets, the result obtained are 170 clusters with a clustering speed of 0.002018130s, while in the neutral datasets, the result obtained are 151 clusters with a clustering speed of 0.001530701s.

Keywords: Online Transportation, Instagram Social Media, Clustering, Single Pass Clustering

1. Pendahuluan

1.1. Latar belakang

Kepadatan penduduk menjadi salah satu problematika di suatu negara termasuk Indonesia. Padatnya penduduk di Indonesia menyebabkan kemacetan yang dapat menimbulkan kerugian waktu, tenaga, bahkan material. Karena hal itu banyak masyarakat yang beralih menggunakan transportasi umum. Banyaknya masyarakat yang beralih ke moda transportasi umum mengakibatkan kepadatan antrean saat ingin menaiki transportasi umum dan saat sudah berada di dalam transportasi umum tersebut. Karena kemacetan dan kepadatan pada transportasi umum tersebut muncul lah moda transportasi *online*.

Dengan adanya transportasi *online* masyarakat dapat mengefisiensikan waktu, tenaga, bahkan material. Penyedia jasa transportasi *online* juga melakukan perkembangan dengan menambahkan berbagai fitur seperti pengiriman makanan, pengiriman barang, jasa pembersih rumah, dan masih banyak lainnya seperti yang dilakukan oleh Go-Jek Indonesia dan Grab Indonesia.

Penyedia jasa transportasi *online* juga tak luput dari berbagai komentar atau penilaian dari penggunanya, baik itu komentar positif, negatif atau netral. Salah satu media yang digunakan pengguna

untuk berkomentar yaitu sosial media Instagram. Pada penelitian – penelitian sebelumnya seperti pada penelitian [1] dan [2] sudah dilakukan klasifikasi terhadap data sentimen pengguna transportasi *online* yang bersumber dari sosial media instagram Go-Jek Indonesia dan Grab Indonesia. Pada penelitian ini dilakukan proses *clustering* terhadap data sentimen pengguna transportasi *online* yang bersumber dari sosial media Instagram Go-Jek Indonesia dan Grab Indonesia. Penelitian ini diharapkan dapat membantu memudahkan pengguna dalam mengetahui kelebihan dan kekurangan dari penyedia jasa transportasi *online* Go-Jek Indonesia dan Grab Indonesia dengan menampil melalui *website*. Pada penelitian ini proses *clustering* menggunakan algoritma *Single Pass Clustering* yang dimana sebelumnya sudah melalui proses TF-IDF.

2. Dasar Teori

2.1. Data Mining

Data mining merupakan penganalisaan data yang berasal dari berbagai pandangan dan diringkas menjadi informasi yang berguna. Sistem *data mining* memiliki komponen penting yaitu *database*, mesin *data mining*, dan modul evaluasi pola [3]. *Data Mining* juga dapat mengidentifikasi dan mengekstraksi informasi dari data menggunakan teknik statistik, matematika kecerdasan buatan, dan *machine learning* untuk mendapatkan hubungan anatara pola dan kecenderungan data dalam jumlah besar [4].

2.2. Pre-processing

Pre-processing merupakan tahapan untuk menyaring kata yang tidak berarti, tidak lengkap, tidak konsisten, dan tidak berstruktur agar menjadi data yang siap untuk diproses ke tahap berikutnya [5].

2.3. Pembobotan Kata

Pembobotan kata merupakan proses mengambil fitur dengan kriterian lebih menonjol untuk dapat diproses ke tahap selanjutnya [6]. Pembobotan Kata dapat memberikan informasi yang sesuai dengan teks, pada penelitian ini digunakan teknik *Term Frequency – Inverse Document Frequency* (TF-IDF) sebagai pembobotan kata. TF-IDF merupakan teknik uang digunakan untuk pengambilan istilah yang paling relevan dengan data dan menghilangkan istilah paling umum, dan akan merubah dokumen menjadi lebih terstruktur. Nilai TF-IDF akan semakin tinggi seiring dengan kemunculannya[7]. Berikut merupakan rumus perhitungan TF-IDF [8] :

$$w_{ij} = tf_{ij} \times id_{fj}$$

$$w_{ij} = tf_{ij} \times log(D / d_{fj}) \quad (1)$$

Dimana:

W_{ij} : Bobot term (tj) terhadap dokumen (di).

tf_{ij} : Jumlah kemunculan term (tj) dalam dokumen (di)

D : Jumlah semua dokumen yang ada dalam *database*.

D_{fj} : Jumlah dokumen yang mengandung term (tj).

2.4. Clustering

Clustering merupakan suatu proses pengelompokan data yang dibagi ke dalam beberapa kelompok atau *cluster* berdasarkan kemiripan antar data yang ada di dalam *cluster* namun memiliki perbedaan dengan *cluster* lainnya [9]. Pada umumnya terdapat tiga proses pada *clustering* yaitu ekstraksi dan pemilihan ciri/fitur, kesamaan antar pola, dan pengelompokan [10].

2.5. Single Pass Clustering

Single Pass Clustering merupakan pengelompokan data dengan mengelompokan data satu demi satu, pembentukan *cluster* atau kelompok dilakukan dengan pengevaluasian kesamaan data yang dimasukan ke dalam *cluster* dan dilakukan evaluasi kesamaan antar *cluster* yang ada [11]. Dimana dokumen pertama yang masuk ke proses *clustering* akan dijadikan menjadi *cluster* pertama. Yang kemudian akan dilakukan perhitungan dengan setiap wakil dari masing – masing *cluster* untuk didaptkan hasil yang kemudian akan dibandingkan dengan nilai *threshold*, jika nilai perhitungan antar dokumen melebihi nilai *threshold* maka dokumen tersebut akan menjadi anggota dari *cluster* yang bersesuaian. Apabila nilai tidak melebihi nilai *threshold* maka dokemen akan dijadikan *cluster* baru, begitupun seterusnya hingga masukan dokumen sudah tidak ada lagi.

2.6. Euclidean Distance

Euclidean Distance digunakan untuk menghitung jarak antar dua titik atau *variable*. Metode ini mudah dan lebih efisien untuk waktu, dan prosesnya yang cepat [12]. Rumus jarak *Euclidean Distance*, yaitu:

$$d = \sqrt{(x2 - x1)2 + (y2 - y1)2} \quad (2)$$

Dimana:

(x1,y1) : Koordinat satu titik.

(x2,y2) : Koordinat titik lainnya.

d : Jarak antara (x1,y1) dan (x2,y2).

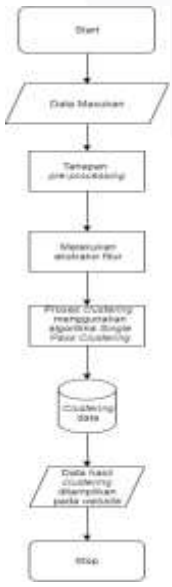
3. Pembahasan

3.1. Desain Sitem

Gambar 1. Desain Sistem

Pada Gambar 1 menjelaskan perancangan sistem pada penelitian ini yang merupakan pengelompokan topik yang dimana, data masukan berupa sentimen dari kolom komentar media sosial Instagram yang telah dilakukan klasifikasi pada penelitian sebelumnya. Selanjutnya dilakukan *pre-processing* yaitu *tokenizing, stopwords, filtering,* dan *stemming*. Kemudian dilakukan proses pembobotan kata dengan TF-IDF. Setelah itu dlakukan proses *clustering* untuk mengelompokan kata berdasarkan sentiment, dan kemudian ditampilkan pada *website*.

3.2. Perancangan Sistem



Gambar 2. *Flowchart Diagram*

Penelitian ini dikerjakan berdasarkan diagram alir pada Gambar 2. Dimana proses pengerjaan dimulai dari data masukan yang telah diklasifikasi pada penelitian sebelumnya, kemudian akan dilakukan tahapan *pre-processing*, kemudian melakukan tahapan pembobotan kata menggunakan metode TF-IDF, setelah dilakukan proses pembobotan kata kemudian melakukan proses *clustering*, yang dimana hasil dari proses *clustering* akan ditampilkan pada *website*.

3.3. Tahapan Sistem

Pada tahapan sistem penelitian ini dijelaskan tahapan *pre-processing* yang dimana terdapat 4 tahapan yaitu *tokenizing, stopwords, filtering,* dan stemming.

1. *Tokenizing* merupakan proses pemisahan atau pemotongan kata berdasarkan kata penyusunnya.

Tabel 1 Proses *Tokenizing*

Data Masukan	Data Keluaran
gojek mmg kasih pelayanan terbaik ke pelanggan keren bgt	“gojek” “mmg” “kasih” “pelayanan” “terbaik” “ke” “pelanggan” “keren” “bgt”

2. *Stopwords* merupakan proses penghapusan kata umum yang sering muncul yang tidak memiliki makna.

Tabel 2 Proses *stopwords*

Data Masukan	Data Keluaran
“gojek” “mmg” “kasih” “pelayanan” “terbaik” “ke” “pelanggan” “keren” “bgt”	gojek kasih pelayanan terbaik pelanggan keren

3. *Filtering* merupakan proses pengambilan kata dengan mengeluarkan kata yang tidak penting dan menyimpan kata yang penting.

Tabel 3 Proses *filtering*

Data Masukan	Data Keluaran
“gojek” “mmg” “kasih” “pelayanan” “terbaik” “ke” “pelanggan” “keren” “bgt”	gojek kasih pelayanan terbaik pelanggan keren

4. *Stemming* merupakan proses memperkecil jumlah indeks yang berbeda dan mengembalikannya ke bentuk dasarnya [13].

Tabel 4 Proses *Stemming*

Data Masukan	Data Keluaran
gojek kasih pelayanan terbaik pelanggan keren	gojek kasih pelayanan terbaik pelanggan keren

4. Pengujian

4.1. Skenario Pengujian

Skenario pengujian yang dilakukan pada penelitian ini dilakukan dengan pengujian terhadap algoritma dan pengujian terhadap sistem. Berikut jenis pengujian untuk pengujian sistem dan pengujian algoritma:

1. Pengujian *Alpha* dilakukan oleh pihak internal untuk menguji fitur – fitur pada *website* dengan metode *black box*, untuk mengetahui fungsi dari fitur -fitur yang tersedia sudah berfungsi dengan baik atau belum.
2. Pengujian *Beta* dilakkan dengan membuat kuisioner yang nantinya akan diberikan kepada responden pihak eksternal untuk menguji pengimplementasian sistem.
3. Pengujian algoritma *Single Pass Clustering* dilakukan untuk mendapatkan hasil *cluster* terbaik dari setiap *dataset*.

4.2. Pengujian Clustering

Pengujian pada *clustering* dilakukan dengan mengubah nilai *threshold* mulai dari 0.1 hingga 0.9 untuk didapatkan hasil pengelompokan kata yang terbaik dari setiap *dataset*.

1. Pengujian *clustering* pada *dataset* positif

Tabel 5 Pengujian Clustering Dataset Positif

Threshold	Jumlah Klaster	Spend Time
0.1	123	0.001196800s
0.2	109	0.001170357s
0.3	99	0.001163048s
0.4	93	0.001094862s
0.5	85	0.001034744s
0.6	79	0.001031061s
0.7	73	0.001029789s
0.8	72	0.000995462s
0.9	66	0.000970668s

Hasil *clustering* pada *dataset* positif dengan mengubah nilai *threshold* mulai dari 0.1 hingga 0.9 didapat hasil *cluster* terbanyak berjumlah 123 *cluster* dengan kecepatan *clustering* 0.001196800s terdapat pada *threshold* 0.1.

2. Pengujian *clustering* pada *dataset* negatif

Tabel 6 Pengujian Clustering Dataset Negatif

Threshold	Jumlah Klaster	Spend Time
0.1	170	0.002018130s
0.2	145	0.001838675s
0.3	127	0.001723533s
0.4	114	0.001670016s
0.5	107	0.001615996s
0.6	101	0.001592555s
0.7	94	0.001583196s
0.8	85	0.001577404s
0.9	79	0.001476580s

Hasil *clustering* pada *dataset* negatif dengan mengubah nilai *threshold* mulai dari 0.1 hingga 0.9 didapat hasil *cluster* terbanyak berjumlah 170 *cluster* dengan kecepatan *clustering* 0.002018130s terdapat pada *threshold* 0.1.

3. Pengujian *clustering* pada *dataset* netral

Tabel 7 Pengujian Clustering Dataset Netral

Threshold	Jumlah Klaster	Spend Time
0.1	151	0.001530701s
0.2	130	0.001461693s
0.3	118	0.001399318s
0.4	111	0.001350974s
0.5	97	0.001299173s
0.6	93	0.001335058s
0.7	88	0.001287109s
0.8	82	0.001233701s
0.9	78	0.001206288s

Hasil *clustering* pada *dataset* netral dengan mengubah nilai *threshold* mulai dari 0.1 hingga 0.9 didapat hasil *cluster* terbanyak berjumlah 151 *cluster* dengan kecepatan *clustering* 0.001530701s terdapat pada *threshold* 0.1.

4.3. Hasil Pengujian

Dari hasil pengujian yang dapat dilihat pada Tabel 5, Tabel 6, dan Tabel 7, dengan menggunakan tiga *dataset* yaitu *dataset* positif, *dataset* negatif, dan *dataset* netral didapatkan hasil uji bahwa pada nilai *threshold* 0.1 didapatkan hasil *cluster* yang lebih banyak dibandingkan dengan nilai *threshold* berikutnya. Namun pada nilai *threshold* 0.1 juga penangkapan data untuk dijadikan *cluster* membutuhkan kecepatan penangkapan yang lebih lama dibandingkan nilai *threshold* berikutnya. Didapatkan bahwa semakin besar nilai *threshold* maka akan semakin cepat dalam menangkap data

untuk dijadikan *cluster* namun jumlah *cluster* menjadi semakin sedikit.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil penelitian, pengujian serta analisa yang telah dilakukan pada penelitian ini, maka penulis mendapat kesimpulan bahwa :

1. Dari pengujian hasil clustering dengan mengubah nilai threshold dari 0.1 sampai 0.9, didapat bahwa semakin besar nilai threshold maka daya tangkap terhadap pengklasteran lebih cepat dan menjadi lebih sedikit clusternya. Pada threshold 0.1 pada dataset positif didapat hasil cluster sebanyak 123 cluster dengan kecepatan pengklasteran 0.001196800s, sedangkan pada threshold 0.9 didapat hasil cluster sebanyak 66 dengan kecepatan lebih cepat dari threshold 0.1 yaitu 0.000970668s. Pada threshold 0.1 pada dataset negatif didapat hasil cluster sebanyak 170 cluster dengan kecepatan pengklasteran 0.002018130s, sedangkan pada threshold 0.9 didapat hasil cluster sebanyak 79 dengan kecepatan lebih cepat dari threshold 0.1 yaitu 0.001476580s. Pada threshold 0.1 pada dataset netral didapat hasil cluster sebanyak 151 cluster dengan kecepatan pengklasteran 0.001530701s, sedangkan pada threshold 0.9 didapat hasil cluster sebanyak 78 dengan kecepatan lebih cepat dari threshold 0.1 yaitu 0.001287109s.
2. Single Pass Clustering dapat mengklasterkan dokumen – dokumen menjadi sebuah cluster – cluster baru yang berdasarkan nilai *threshold* dan nilai *similarity* antar vektor datanya dengan menggunakan *euclidean distance*.

5.2. Saran

Saran yang bisa penulis usulkan untuk penelitian lebih lanjut yaitu :

1. Dapat menggunakan perhitungan yang lain selain Euclidean distance agar dapat mengasilkan cluster yang lebih baik.
2. Mengembangkang atau menggunakan algoritma lain untuk proses clustering agar menghasilkan hasil cluster yang lebih baik.

Referensi

[1] S. Rohwinasakti, B. Irawan. C. Setianingsih, ”SENTIMENT ANALYSIS ON ONLINE TRANSPORTATION SERVICE USING K-NEAREST NEIGHBOR,” 2020

[2] D.S. Ashari, B. Irawan. C. Setianingsih, ”SENTIMENT ANALYSIS ON ONLINE TRANSPORTATION SERVICE USING CNN (CONVULUTIONAL NEURAL NETWORK),” 2020

[3] R. I. K. Rao, “Data Mining and Clustering Techniques,” no. November, 2014.

[4] M. K. Siregar, Amril Mutoi, S.Kom. and M. K. Puspabhuana, Adam, S.Kom., *DATA MINING: Pengolahan Data Menjadi Informasi dengan RapidMiner*. CV Kekata Group.

[5] S. Srivastava, “Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining,” *Int. J. Comput. Appl.*, vol. 88, no. 10, pp. 26–29, 2014, doi: 10.5120/15389-3809.

[6] H. (National U. of S. Liu, H. (Osaka U. Motoda, R. Setiono, and Z. Zhao, “Feature Selection : An Ever Evolving Frontier in Data Mining,” *J. Mach. Learn. Res. Work. Conf. Proc. 10 Fourth Work. Featur. Sel. Data Min.*, pp. 4–13, 2010.

[7] P. Bafna, D. Pramod, and A. Vaidya, “Document clustering: TF-IDF approach,” *Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016*, pp. 61–66, 2016, doi: 10.1109/ICEEOT.2016.7754750.

[8] Informatikalogi, “Pembobotan Kata atau Term Weighting TF-IDF,” *informatikalogi.com*. https://informatikalogi.com/term-weighting-tf-idf/ (accessed Nov. 25, 2020).

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

[10] I. Werdiningsih, B. Nuqoba, and Muhammadun, “Data Mining Menggunakan Android, Weka, Dan Spss”. Airlangga University Press, 2020.

[11] F. Herny, Z. Eri, “ALGORITMA SINGLE PASS CLUSTERING UNTUK KLASTERING HALAMAN WEB,” Unisbank Semarang.

[12] Pamungkas, Canggih Ajika, “Aplikasi Penghitung Jarak Koordinat Berdasarkan Latitude dan Longitude Dengan Metode Euclidean Distance Dan Metode Haversine,” Jurnal Informa Politeknik

Indonesia Surakarta., Vol. 5 Nomor 2 Tahun 2019.

[15] Informatikalogi, “Text Preprocessing,” informatikalogi.com. https://informatikalogi.com/text-preprocessing/ (accessed Dec. 01, 2020).