CLUSTERING PADA DATA SENTIMEN TRANSPORTASI ONLINE MENGGUNAKAN ALGORITMA DBSCAN

CLUSTERING ON SENTIMENT DATA ONLINE TRANSPORTATION USING DBSCAN ALGORITHM

Firdi Setiawan[1], Fairuz azmi[2], Casi Setianingsih[3]

[1,2,3] Universitas Telkom , Bandung
firdisetiawan@student.telkomuniversity.ac.id[1], worldliner@telkomuniversity.ac.id[2], setiacasie@telkomuniversity.ac.id[3]

Abstrak

Penelitian ini dilakukan pengelompokan sentimen pada masing –masing data sentimen positif negatif, dan netral menggunakan algoritma DBSCAN (Density-Based Spatial Clustering of Applications With Noise), Tujuan utama dari clustering ini untuk mengelompokkan opini masyarakat yang berdasar pada kesamaan karakteristik atau makna dalam penulisan di antara opini-opini tersebut untuk menentukan positif, negatif, dan netral berdasarkan komentar pada media sosial instagram. Dengan melakukan tahapan preprocessing seperti tokenize, stopword, dan stemming, kemudian dilakukan pembobotan kata dengan menggunakan TF- IDF untuk dapat melakukan pengelompokan opini. Dari hasil Clustering didapatkan hasil dari pengujian dataset positif, negatif, dan netral masing masing diuji coba dengan range nilai min sampel dari 10-50 dan nilai epsilon dari 0,1-1,0 dengan menghasilkan nilai silhouette coefficientnya berbeda beda. Namun untuk nilai terbaik dari ketiga dataset didapatkan pada nilai inputan eps=1,0 dan inputan nilai min sampel = 10, untuk hasil dataset positif nilai silhouette coefficient-nya adalah 0.7800973549904059, untuk hasil dataset netral nilai silhouette coefficient-nya adalah 0.7526159947007542, untuk hasil dataset negatif nilai silhouette coefficient-nya adalah 0.8047251594403672. Kemudian visualisasi data hasil clustering topik tersebut akan ditunjukkan pada perangkat lunak berbasis web yang juga dirancang pada penelitian ini.

Kata kunci : Clustering, Preprocessing, Silhouette Coefficient

Abstract

In this research, sentiment grouping will be carried out on each positive, negative, and neutral sentiment data using the DBSCAN (Density-Based Spatial Clustering of Applications With Noise) algorithm. The main purpose of this clustering is to classify public opinion based on similar characteristics or meaning in writing on between these opinions to determine positive, negative, and neutral based on comments on Instagram social media. By doing preprocessing stages such as tokenize, stopword,and stemming, then word weighting is carried out using TF-IDF to be able to group opinions. From the Clustering results, it is obtained that the positive, negative, and neutral datasets were tested with a minimum sample value range of 10-50 and an eps value of 0.1-1.0 by producing different silhouette values. However, the best value from the third dataset is obtained at the input value of eps = 1.0 and the inputmin sample value = 10, for the positive dataset the silhouette coefficient value is 0.7800973549904059, for the neutral dataset the silhouette coefficient value is 0.7526159947007542, for the negative dataset the silhouette value is the coefficientis 0.8047251594403672. Then the data visualization of the results of the topic grouping will be shown on a web-based software which is also designed in this final project research.

Keywords: Clustering, Preprocessing, Silhouette Coefficient.

1. Pendahuluan

Indonesia merupakan salah satu negara di Asia dengan jumlah pengguna jasa transportasi *online* terbesar, menurut laporan We Are Social 2020, saat ini terdapat sebanyak 21,7 juta orang di Indonesia yang menggunakan layanan transportasi *online* [1]. Memang beberapa tahun belakangan ini transportasi*online* sangat populer bagi masyarakat Indonesia, selain dapat memudahkan kita

untuk bepergian, layanan jasa transportasi *online* ini juga menyediakan berbagaikemudahan lain seperti layanan pembayaran, layanan pembersihan rumah, hingga pemesanan makanan dan juga akses penggunaan jasanya mudah dilakukan cukup dengan menggunakan *Smartphone* yang terhubung dengan *internet*, alat yang tentunya biasa masyarakat gunakan sehari-hari.

Dengan berkembangnya penggunaan layanan transportasi *online*, masyarakat semakin tidak asing lagi mendengarnya karena banyak masyarakat yang membicarakan tentang pengalaman menggunakan layanan transportasi *online*. Masyarakat membicarakan dan memberikan opini seputar pengalaman menggunakan transportasi *online* melalui berbagai media salah satunya adalah media sosial Instagram. Menurut data penelitian We Are Social, Instagram beradapada urutan keempat terbesar di Indonesia pada survei *platforms* media sosial yang paling aktif. Pengguna Instagram di Indonesia sebanyak 79% dari populasi [2].

Beragam opini yang diaspirasikan oleh masyarakat terhadap layanan jasa transportasi *online*, seperti pelayanan, aplikasi, ataupun sikap pengemudi. Pada Instagram, layanan transportasi *online* tentunya memiliki akun resmi untuk memberikan informasi terbaru tentang layanan maupun mengumpulkan komentar-komentar dari masyarakat atau pelanggan. Apabila diteliti lebih lanjut terhadap kumpulan komentar tersebut, maka akan didapatkan sebuah sentimen yang apabila dikumpulkan akan mendapatkan kesimpulan sentimen masyarakat terhadap jasa transportasi *online*.

Pada penelitian-penelitian sebelumnya seperti pada penelitian [3] dan [4] sudah dilakukan klasifikasi mengenai sistem analisis sentimen dalam menilai sentimen yang bersumber dari komentar pada media sosial penyedia jasa transportasi *online.* Penelitian ini dimaksudkan untuk melakukan *clustering* data sentimen masyarakat terhadap penyedia layanan transportasi *online* Indonesia pada media sosial Instagram agar membantu masyarakat Indonesia yang menggunakan jasa layanan transportasi *online* dengan menampilkan masing masing keunggulan dan kekurangan layanan produk dari penyedia layanan transportasi *online* dari hasil *clustering* sentimen. *Clustering* data sentimen ini dilakukan dengan menggunakan algoritma DBSCAN (*Density-Based Spatial Clustering ofApplications With Noise*).

2. Dasar Teori

2.1 Text Mining

Text Mining adalah suatu bidang untuk pengolahan data dalam bentuk teks tidak terstruktur (*unsupervised*) dan memiliki tujuan untuk menganalisis dan memproses teks menjadi suatu informasi yang menggunakan beberapa teknik standar yaitu klasifikasi teks, pengelompokan teks, pembuatan ontologi dan taksonomi, peringkasan dokumen dan analisis *korpus laten* [5].

2.2 Clustering

Clustering merupakan suatu metode yang biasanya digunakan untuk menganalisis suatu data yang biasa dapat di kelompokkan dengan menggunakan pengenalan pola dan menjadi salah satu metode yang dapat digunakan pada *data mining*. *Clustering* ini bersifat mandiri yang biasanya disebut dengan *unsupervised learning* karena metode ini tidak membutuhkan data latih yang sudah diketahui kelasnya, dan *clustering* mengelompokkan data hanya berdasar pada fitur yang ada pada data tersebut [6]. Tujuan utama dari proses *clustering* yaitu membagi sekumpulan data menjadi sebuah kelompok, data dikelompokkan berdasarkan prinsip untuk mengumpulkan data data yang terdapat dalam satu *cluster* yang sama dan juga ketidaksamaan dengan data yang ada pada *cluster* lain.[7]

2.3 Pembobotan Kata

Pembobotan kata merupakan metode untuk meringkas suatu dokumen yang berdasar pada bobot kata. Contoh pembobotan kalimatyang berdasarkan kata adalah *word frequency* (WF) dan TF-IDF. TF- IDF merupakan suatu metode pembobotan frekuensi kata berdasarkan intenstas kemunculannya. pada suatu dokumen teks. Pembobotan ini dapat dilakukan menggunakan persamaan berikut [8] :

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (1)$$

Dimana :

1. Nilai tfidf (t, d,D)i diketahui merupakan perhitungan frekuensi *term* dari dokumen teks.
2. Nilai d,D merupakan jumlah keseluruhan dokumen.
3. t merupakan banyaknya *term*.

4. tf(t, d) merupakan term frekuensi dari suatu dokumen teks.
5. idf (t, D) merupakan *inverse* frekuensi dokumen

2.4 Pre-Processing

Pre-processing adalah *processing* atau praproses data merupakan tahapan proses untuk mempersiapkan kumpulan data mentah yang belum dilakukan tahapan proses lain. Tahapan *pre-processing* ini pada umumnya dilakukan dengancara mengeliminasi sebuah data yang tidak sesuai dengan format atau mengubah data menjadi bentuk yang lebih mudah untuk diproses oleh sistem [9]. *Pre-processing* sangat dibutuhkan dan berguna dalam melakukan analisis sentimen, terutama untuk memproses komentar pada media sosial yang sebagian besar berisi kata-kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar [10].

2.5 DBSCAN (*Density-Based Spatial Clustering of Application with Noise*)

Density-Based Spatial Clustering of Application with Noise (DBSCAN) sebuah metode pengelompokan berbasis kepadatan yang tidak hanya dapat menemukan *cluster* bentuk sewenang-wenang dan menanganititik-titik kebisingan, tetapi juga dapat mendeteksi jumlah *cluster* secara alami [11]. DBSCAN juga merupakan metode *clustering* yang membangun area berdasarkan densitas yang terkoneksi (*density connected*). DBSCAN adalah jenis *clustering* yang mengkategorikan kerapatan tinggi dari suatu daerah menjadi *cluster* yang ditemukan dalam bentuk bebas dengan memanfaatkan *noise* [12].

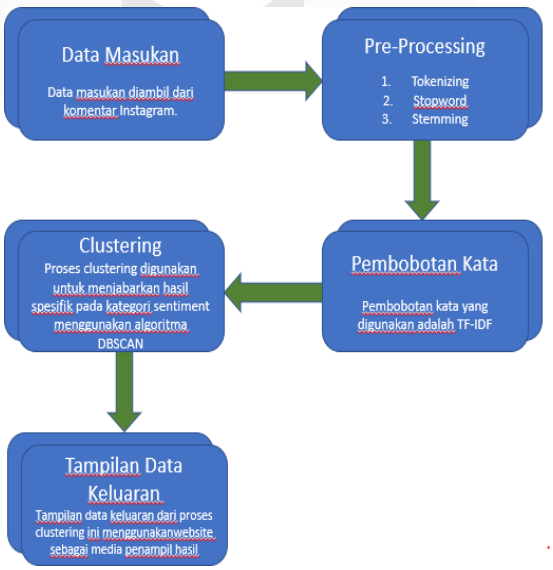
2.6 Silhouette Coefficient

Metode *silhouette coefficient* merupakan gabungan dari dua metode yaitu metode *cohesion* yang berfungsi untuk mengukur seberapa dekat relasi antara objek dalam sebuah *cluster*, dan metode *separation* yang berfungsi untuk mengukur seberapa jauh sebuah cluster terpisah dengan cluster lain [13].

3. Perancangan Sistem

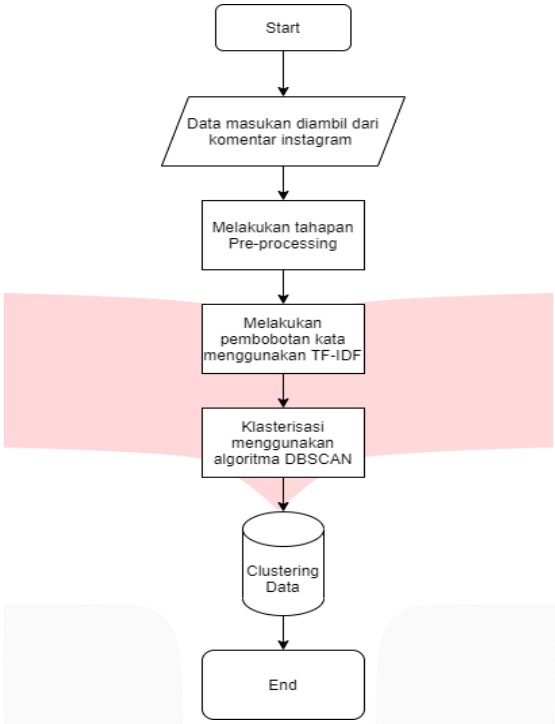
3.1. Desain Sistem

Desain sistem pada Gambar 1 adalah desain sistem untuk keseluruhan proses dari *clustering*, dimulai dari pengambilan data pada komentar dari akun Instagram gojek dan grab yang sudah di proses sampai pada tahap klasifikasi yang dilakukan pada penelitian sebelumnya. Kemudian data dilakukan *pre-processing* dengan menggunakan *tokenizing*, *stopword*, dan *stemming*. Selanjutnya data diproses dengan pembobotan kata menggunakan metode TF-IDF dimana kata diubah menjadi bentuk *array*, setelah itu dilakukan proses *clustering* untuk mengelompokkan kata berdasarkan sentimen masyarakat yang kemudian ditampilkan kedalam *website*



Gambar 1. Desain Sistem

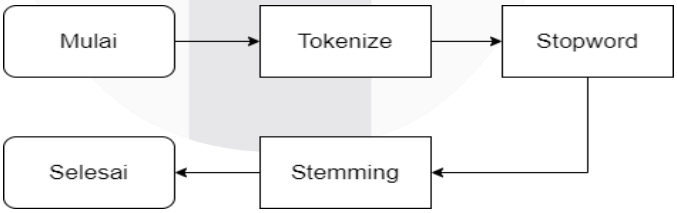
Pengerjaan penelitian dilakukan berdasarkan rancangan diagram alir yang sudah dibuat seperti pada Gambar 2. Pada diagram alir telah dipaparkan langkah-langkah pengerjaan mulai dari melakukan pengumpulan data, kemudian melakukan proses *pre-processing*, dilanjutkan dengan pembobotan kata menggunakan metode pembobotan TF-IDF, yang selanjutnya dilakukan proses *clustering* untuk tiap dataset positif, negatif, dan netral, kemudian data hasil *clustering* disimpan kedalam database yang kemudian ditampilkan didalam laman website.



Gambar 2. *Flowchart* Sistem

3.2. Desain Proses Tahapan Sistem

Dalam penelitian ini proses tahapan sistem menjelaskan detail kerja dari *pre-processing* dengan tiga tahapan proses, yaitu *tokenizing*, *stopword*, dan *stemming*.



Gambar 3. Tahapan Pre-Processing

1. Tokenization

Tahap *tokenization* dilakukan untuk memisahkan kumpulan karakter menjadi sebuah satuan kata.

Tabel 1 Proses *Tokenization*

Data Masukan	Data Keluaran
abang drivernya membahayakan penumpang	“abang”, “drivernya”, “membahayakan”, “penumpang”

2. Stopword

Tahapan untuk pemilahan kata-kata dari hasil proses *tokenizing* dengan menghapus kata hubung seperti kata “dan”, “atau”.

Tabel 2 Proses *Stopword*

Data Masukan	Data Keluaran
Abang drivernya membahayakan penumpang	“abang”, “drivernya”, “membahayakan”,“penumpang”

3. Stemming

Tahap *stemming* ini bertujuan untuk menghapus semua kata imbuhan agar teks hanya terdiri dari kata dasar.

Tabel 3 Tahap *Stemming*

Data Masukan	Data Keluaran
abang drivernya membahayakan penumpang	abang driver bahaya tumpang

3.3. Skenario Pengujian

Skenario performansi yang akan dilakukan adalah pengujian algoritma dan pengujian sistem. Pengujian algoritma untuk menguji atau mengukur performansi dari sisi algoritma *DBSCAN*. Pengujian sistem dilakukan untuk menguji fungsionalitas sistem ketika selesai diimplementasikan.

Jenis pengujian sistem dan algoritma yang akan dilakukan pada penelitian iniadalah sebagai berikut:

a. Pengujian Alpha : Pengujian ini dilakukan untuk menguji hasilpengembangan perangkat lunak dalam mengukur apakah sistem dikembangkan dengan baik tanpa ada error atau kesalahan.
b. Pengujian Beta : Pengujian ini akan dilakukan untuk melihat tanggapan dari sisi pengguna mengenai implementasi sistem.
c. Pengujian Algoritma DBSCAN : Pengujian ini dilakukan untuk mendapatkan hasil *cluster* terbaik dari setiap *dataset*

4. Pengujian

4.1. Pengujian *Clustering*

Pada penelitian ini telah dilakukan pengujian dengan jumlah sebanyak 30 pengujian. dengan 3 dataset positif, negatif, dan netral dengan masing masing dilakukan 10 kali pengujian dengan perubahan nilai eps dari 0.1-1.0 dan nilai minsampel/minpts dari 10. Pada tabel 4, 5, dan 6 akan ditampilkan contoh pengujian *clustering* dengan hasil terbaik dan masing-masing contoh 1 tabel untuk setiap dataset.

a. Hasil Pengujian Dataset Positif

Pada contoh pengujian dataset positif pada table 4 dilakukan sebanyak 10 kali pengujian dengan memasukkan nilai minimum sampel 10 dan nilai *epsilon* dari 0.1-1.0 di setiap nilai minimum sampel dan menghasilkan *silhouette coefficient* terbaik sebesar 0.7800973549904059 terletak pada uji coba nilai minimum sampel 10 dan *epsilon* 1.0

Tabel 4 Pengujian positif min sampel 10

Pengujian Dengan Nilai Minsampel 10		
Eps=0,1	Estimasi Cluster = 6	SC = -0.42875137974626043
Eps=0,2	Estimasi Cluster = 1	SC = 0.47858336866886214
Eps=0,3	Estimasi Cluster = 2	SC = 0.3597240477290904
Eps=0,4	Estimasi Cluster = 2	SC = 0.550174650347391
Eps=0,5	Estimasi Cluster = 1	SC = 0.7076156591444477
Eps=0,6	Estimasi Cluster = 1	SC = 0.7578313081834464
Eps=0,7	Estimasi Cluster = 1	SC = 0.763933958616549
Eps=0,8	Estimasi Cluster = 1	SC = 0.763933958616549
Eps=0,9	Estimasi Cluster = 1	SC = 0.7800973549904059
Eps=1,0	Estimasi Cluster = 1	SC = 0.7800973549904059

b. Hasil Pengujian Dataset Netral

Pada contoh pengujian dataset positif pada table 5 dilakukan sebanyak 10 kali pengujian dengan memasukkan nilai minimum sampel 10 dan nilai *epsilon* dari 0.1-1.0 di setiap nilai minimum sampel dan menghasilkan *silhouette coefficient* terbaik sebesar 0.7526159947007542 terletak pada uji coba nilai minimum sampel 10 dan *epsilon* 1.0

Tabel 5 Pengujian netral min sampel 10

Pengujian Dengan Nilai Minsampel 10		
Eps=0,1	Estimasi Cluster = 9	SC = -0.4841487440906893
Eps=0,2	Estimasi Cluster = 2	SC = 0.33125155336219914
Eps=0,3	Estimasi Cluster = 3	SC = 0.4626537393391486
Eps=0,4	Estimasi Cluster = 1	SC = 0.6813467455580137
Eps=0,5	Estimasi Cluster = 1	SC = 0.7084296273051851
Eps=0,6	Estimasi Cluster = 1	SC = 0.7280870961270355
Eps=0,7	Estimasi Cluster = 1	SC = 0.7283276086163007
Eps=0,8	Estimasi Cluster = 1	SC = 0.744609927234066
Eps=0,9	Estimasi Cluster = 1	SC = 0.7521995798476576
Eps=1,0	Estimasi Cluster = 1	SC = 0.7526159947007542

c. Hasil Pengujian Dataset Negatif

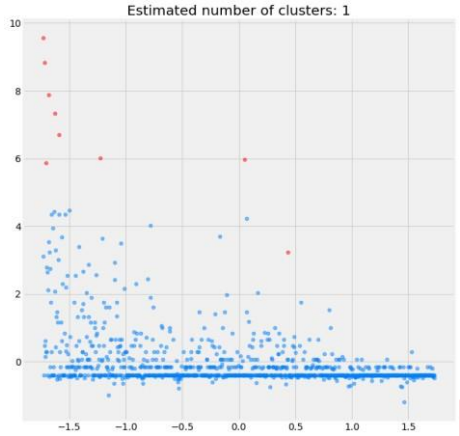
Pada contoh pengujian dataset positif pada tabel 6 dilakukan sebanyak 10 kali pengujian dengan memasukkan nilai minimum sampel 10 dan nilai *epsilon* dari 0.1-1.0 di setiap nilai minimum sampel dan menghasilkan *silhouette coefficient* terbaik sebesar 0.8047251594403672 terletak pada uji coba nilai minimum sampel 10 dan *epsilon* 1.0

Tabel 6 Pengujian negatif min sampel 10

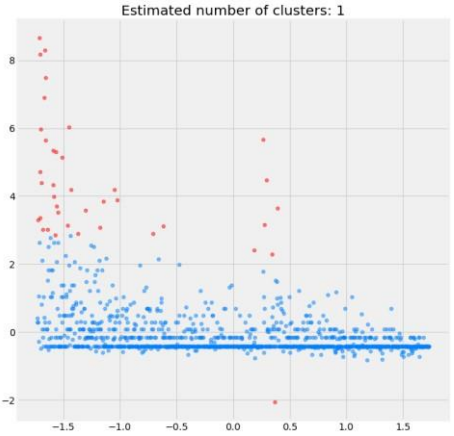
Pengujian Dengan Nilai Minsampel 10		
Eps=0,1	Estimasi Cluster = 3	SC = 0.3349580984331593
Eps=0,2	Estimasi Cluster = 1	SC = 0.5802203972607828
Eps=0,3	Estimasi Cluster = 3	SC = 0.4833464126283004 7
Eps=0,4	Estimasi Cluster = 1	SC = 0.7054226042494941
Eps=0,5	Estimasi Cluster = 1	SC = 0.717566061089248
Eps=0,6	Estimasi Cluster = 1	SC = 0.6921526450315381
Eps=0,7	Estimasi Cluster = 1	SC = 0.7564575448658728
Eps=0,8	Estimasi Cluster = 1	SC = 0.771145726624113
Eps=0,9	Estimasi Cluster = 1	SC = 0.771145726624113
Eps=1,0	Estimasi Cluster = 1	SC = 0.8047251594403672

4.2. Hasil Pengujian

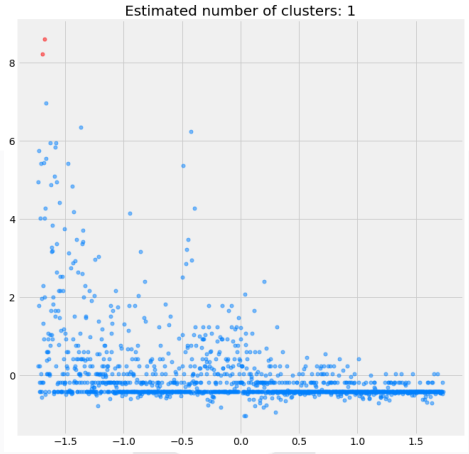
Dari ketiga dataset diatas positif, negatif, dan netral masing masing diuji coba dengan *range* nilai minimum sampel = 10 dan nilai eps dari 0,1-1,0 dengan total 30 pengujian dan dari ketiga dataset yang diuji menghasilkan nilai *silhouette coefficient* yang berbeda beda. Namun untuk nilai terbaik dari ketiga dataset didapatkan pada nilai inputan eps=1,0 dan inputan nilai min sampel = 10, untuk hasil pengujian terbaik dari dataset positif nilai *silhouette coefficient*-nya adalah 0.7800973549904059, untuk hasil dataset netral 0.7526159947007542, untuk hasil dataset negatif 0.8047251594403672.



Gambar 4. Plot hasil terbaik dataset positif



Gambar 5. Plot hasil terbaik dataset netral



Gambar 6. Plot hasil terbaik dataset negatif

5. Kesimpulan dan Saran

5.1 Kesimpulan

Setelah dilakukan penelitian ini, kesimpulan yang dapat diambil pada TugasAkhir ini adalah sebagai berikut:

1. Sistem berhasil melakukan *clustering* komentar pengguna berupa sentimen Positif, Negatif dan Netral dengan algoritma *DBSCAN* didapat nilai *silhouette coefficient* terbaik dari ketiga dataset adalah pada dataset sentimen negatif yaitu sebesar 0.8047251594403672 pada nilai inputan min sampel=10 dan eps=100
2. Pada Dataset positif nilai *cluster* terbaik terdapat pada nilai inputan min sampel=10 dan eps=1.0 dan 0.9 dengan nilai *silhouette coefficient* 0.7800973549904059, pada dataset netral nilai *cluster* terbaik sama dengan dataset positif yaitu dengan nilai 0.7526159947007542, dan pada dataset negatif nilai *cluster* terbaik terdapat pada nilai inputan min sampel = 10 dan eps = 1.0 dengan nilai 0.8047251594403672

5.2 Saran

Hasil penelitian, pengujian dan analisa telah dilakukan pada tugas akhir ini, maka saran yang dapat diusulkan untuk penelitian lebih lanjut yaitu:

1. Dapat menggunakan metode klustering yang lain yang lebih cocok untuk dipadukan dengan metode lain selain TF-IDF untuk mendapatkan nilai *silhouette coefficient* terbaik diatas 0.9 .
2. Menambahkan produk layanan lain agar lebih bervariasi pada saat melakukan perbandingan layanan.

REFERENSI

[1] Y. Astutik, “21,7 Juta Masyarakat Indonesia Pakai Transportasi Online.” CNBC INDONESIA, 17Maret 2020. [Online]. Available: https://www.cnbcindonesia.com/tech/20200317150135-37-145529/217- juta-masyarakat-indonesia-pakai-transportasi-online. [Accessed 28November2020].

[2] S. KEMP, "Hootsuite (We are Social): DIGITAL 2019 INDONESIA,"DATAREPORTAL, 31 Januari 2019. [Online]. Available: https://datareportal.com/reports/digital-2019-indonesia. [Accessed 28November 2020].

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

[4] D. S. Ashari, B. Irawan, and C. Setianingsih, “Sentiment Analysis on Online Transportation Service ’ S Using Cnn (Convolutional Neural Network) Method,” 2020.

[5] Tania, A. S. Restanti and e. al., "Dunia Berubah dengan Keterlibatan Dunia,"in *Media Sosial, Identitas, Transformasi, dan Tantangannya*, Intrans Publishing Group, 2020, p. 61.

[6] D Meyer, K. Hornik and I. Feinerer, "Text Mining Infrastructurein R,"*Journal of Statistical Software*, vol. 25, no. 5, 2008.

[7] L. Ma, L. Gu, B. Li, S. Qiao and J. Wang, "G-DBSCAN: An Improved DBSCAN Clustering Method," *Advanced Science and Technology Letters*,vol. 74, no. Asea, pp. 23-28, 2014

[8] C. Z. Charu and C. Anggarwal, "A SURVEY OF TEXT CLUSTERING ALGORITHM, in Mining Text Data," no. Hawthrone and Urbana, pp. 77-128,2012.

[9] A. B. King, Website Optimization, O'Reilly Media, 2008.

[10] V. R and R. P. M., "Pre-processing and post-processing in group–clustermergers," vol. 2735, pp. 2713-2735, 2013.

[11] I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris and I. Antoniou, "Probabilistic density-based estimation of the number of clusters using theDBSCAN-martiangle process," *Pattern Recognition Letters*, vol. 123, pp. 23-30, 2019.

[12] B. Setiyono and I. Mukhlash, "Kajian Algoritma GDBScan , Clarans danCureuntuk Spatial Clustering," vol. 2, no. 2, pp. 117-128, 2005

[13] J. Foer, Moonwalking with Einstein, Kindle Edi. Washington DC, District Columbia, United States: Penguin, 2011