DOCUMENT CLUSTERING TERKAIT HEALTH NEWS PADA TWITTER DATA SET MENGGUNAKAN K-MEANS CLUSTERING

Dimas Yudhistira Wijaya Koesuma[1], Ari Hernawan S.Kom., M.Sc.[2], Raphael Bianco Huwae S.T., M.T.[3]

[1]Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA

Email: dimas.yudhistira888@gmail.com, arihernawan@unram.ac.id, Raphael.bianco.huwae@unram.ac.id

***Abstract*– Clustering is a data analysis method aimed at grouping data with similar characteristics into the same area. Document clustering involves grouping documents based on their own characteristics. The large amount of health document data in Twitter accounts can significantly contribute to slowing down the document search process. The Health News-related dataset on Twitter consists of health-related information from multiple news agency accounts. This dataset was obtained in 2015 using the Twitter API, involving more than 15 leading health news agencies. Each agency has information about each of their tweets. This research was conducted to assess the performance and application of K-Means clustering. Furthermore, the clustering performance was evaluated using the K-Means method with the silhouette coefficient. In this study, the K-Means algorithm was iterated 5 times, resulting in a Sum of Square Error (SSE) ranging from 3288.3146163817337 to 3402.023678158529 and a silhouette coefficient ranging from 0.0027409235213496225 to 0.003479863529015515. Therefore, it can be concluded that the quality of the clustering performed is relatively low.**

Keywoard **– *K-Means, Spark,* Clustering, Silhouette coefficient, *Health News***

I. PENDAHULUAN

Clustering adalah metode analisis data yang tujuannya mengelompokkan data dengan karakteristik yang sama ke suatu wilayah yang sama. *Document clustering* merupakan pengelompokan dokumen berdasakan karakteristik dokumen itu sendiri[1]. Banyaknya jumlah data dokumen kesehatan dalam sebuah akun dari twitter dapat memberi kontribusi besar dalam lambatnya proses pencarian suatu dokumen. Pencarian dokumen yang ada saat ini hanya menampilkan hasil pencarian berurut berdasarkan peringkat kecocokan (*document ranking*). Hal tersebut menyebabkan penemuan data dokumen tidak secara akurat. Data set terkait Health News pada Twitter merupakan informasi terkait kesehatan dari beberapa akun agensi berita. Data set ini di peroleh pada tahun 2015 menggunakan tweet dengan lebih dari 15 agensi berita kesehatan terkemuka. Setiap agensi memiliki informasi pada setiap *tweet*-nya. Informasi yang diberikan mengandung data berupa id *tweet*, tanggal dan waktu, dan *caption* dari *tweet*-nya. Metode yang digunakan dalam penelitian ini adalah K-Means *clustering*. K-Means clustering merupakan metode pengklasteran yang memisahkan data kedalam k kelompok yang berbeda artinya sebelum dilakukan klasterisasi maka perlu menentukan jumlah k yang diinginkan[9]. Metode K-Means merupakan metode clustering yang cukup sederhana dan umum dalam penggunaannya. KMeans seringkali digunakan dalam permasalahan *clustering* dikarenakan mempunyai kemampuan pengelompokkan data dalam jumlah yang cukup besar dan dengan waktu komputasi yang relatif cepat serta efisien[8]. Dalam aplikasi data mining, K-Means Clustering dapat digunakan bersama dengan TF-IDF untuk mengelompokan dokumen berdasarkan topik yang mirip. Prosesnya adalah dengan menghitung nilai TF-IDF untuk setiap kata dalam setiap dokumen, kemudian menggunakan

nilai-nilai tersebut sebagai fitur untuk K 2 Means *Clustering*. Dengan demikian, dokumen yang memiliki kata-kata yang sama akan dianggap sebagai dokumen yang mirip dan akan dikelompokkan ke dalam kategori yang sama.

II. TINJAUAN PUSTAKA

A. *Text Mining*

Teks Mining didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining*, dimana salah satu fungsinya adalah kategorisasi[2]. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi struktur[3]. Text mining sering digunakan untuk mengidentifikasi pola, menemukan hubungan, dan membuat generalisasi dari data teks. Gambar 2.1 menunjukkan bahwa terdapat 5 dokumen dengan kata pencarian cocok yang ditemukan pada setiap dokumen secara berurutan sebanyak 4, 3, 0, 1, dan 0.

B. *Term Frequency – Inverse Document Frequency (TF-IDF)*

TF adalah salah satu metode pembobotan *term* yang paling sederhana, caranya adalah dengan menghitung jumlah kata atau term yang muncul dalam satu dokumen. Setiap *term* t diasumsikan memiliki kepentingan yang proporsional terhadap jumlah kemunculan *term* pada dokumen d. Dengan metode ini, nilai kontribusi (bobot) suatu *term* pada suatu dokumen adalah sama dengan jumlah munculnya *term* tersebut pada dokumen[12].

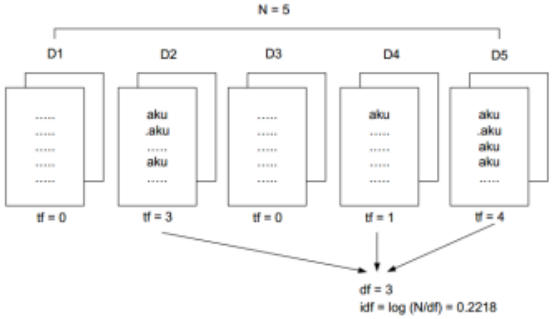
$$TF(d,t) = f(d,t)$$

Dimana f(d, t) adalah frekuensi kemunculan term t pada dokumen d. *Inverse Document Frequency* (IDF) adalah metode perhitungan bobot yang mirip dengan TF hanya saja pada IDF mencari kemunculan *term* pada kumpulan dokumen, berbeda dengan TF yang hanya memerhatikan kemunculan term pada dokumen tersebut. Dalam kata lain IDF memperhatikan jumlah dokumen d yang memiliki kata atau *term* t. Pembobotan ini dilakukan untuk memberikan nilai yang tinggi pada *term* t yang jarang muncul pada kumpulan dokumen d karena term t sangat bernilai. Kepentingan tiap term t diasumsikan memiliki proporsi yang berkebalikan dengan jumlah dokumen d yang mengandung *term* t. Perhitungan nilai *Inverse Document Frequency*[10].

$$IDF(t) = log_{df(t)}$$

Setelah menemukan nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF), maka akan dilakukan pembobotan TF-IDF dengan mengalikan nilai TF dengan nilai IDF. Perhitungan nilai TF-IDF dapat dilihat sebagai berikut.

$$TF.IDF = TF(d,t) * IDF(t)$$



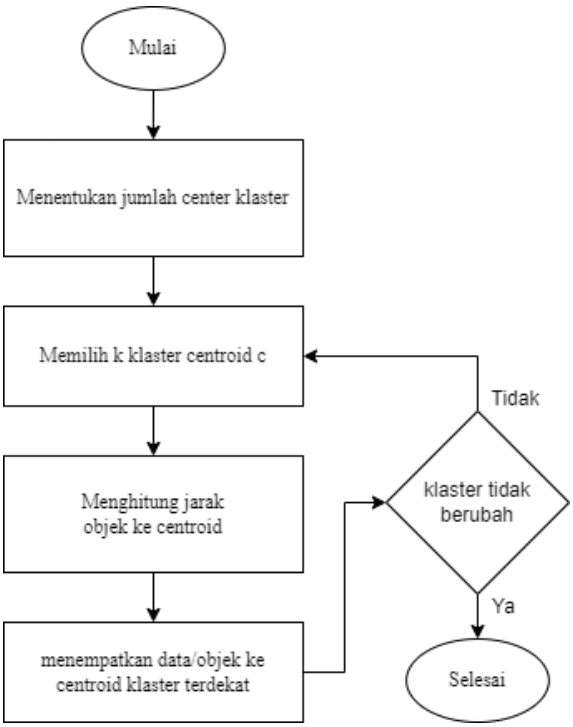
C. *Clustering*

Clustering adalah sebuah teknik pembelajaran mesin yang digunakan untuk mengelompokkan data menjadi kelompok-kelompok yang serupa atau terkait dengan satu sama lain. Kelompok-kelompok ini biasanya disebut sebagai klaster. Proses ini dilakukan tanpa menggunakan label atau informasi tentang kelompok mana data tersebut seharusnya masuk. Tujuan dari *clustering* adalah untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki karakteristik atau kesamaan yang kuat, sehingga data dalam kelompok yang sama lebih mirip satu sama lain daripada data di kelompok lain. *Clustering* dapat digunakan untuk mengidentifikasi pola dan struktur dalam data, serta untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki kesamaan yang kuat[6].

D. *K-Means Clustering*

K-Means merupakan salah satu metode data *clustering* non hierarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lainnya. K-Means adalah metode *clustering* berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini haya bekerja pada atribut numeric. Algoritma K-Means termasuk *partitioning clustering* yang memisahkan data ke k daerah bagian yang terpisah. Algoritma K-Means sangat

terkenal karena kemudahan dan kemampuannya untuk mengcluster data yang besar dan data *outlier* dengan sangat cepat. Dalam algoritma K-Means, setiap data harus termasuk ke *cluster* tertentu dan bisa dimungkinkan bagi setiap data yang termasuk *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke *cluster* lainnya[13].



E. Python

Python adalah bahasa pemrograman tingkat tinggi yang menyediakan kontrol struktur yang fleksibel, tipe data yang sederhana, dan dukungan terhadap berbagai macam paradigma pemrograman, seperti pemrograman berorientasi objek, pemrograman imperatif, dan pemrograman fungsional. Python juga memiliki sejumlah *library* dan pustaka yang luas yang dapat digunakan untuk berbagai keperluan, seperti pemrosesan teks, pemrosesan data numerik, dan pengembangan aplikasi web[5].

F. Apache Spark

Apache Spark adalah "mesin analitik terpadu untuk pemrosesan data skala besar". Spark memiliki banyak fitur, yang membuatnya menjadi pilihan optimal untuk analisis data besar. Misalnya, dalam pemrosesan paralel, aplikasi dapat dengan mudah dibuat menggunakan lebih dari 80 operator tingkat tinggi yang disediakan oleh Spark. Spark juga menawarkan shell interaktif yang ditulis dalam bahasa pemrograman yang berbeda seperti Python, Scala, atau R. Selain itu, Spark mengandung sebagian besar library yang dibutuhkan untuk langkahlangkah analisis data besar.

G. PENELITIAN TERDAHULU

Penelitian yang dilakukan oleh Hafiz Irsyad dan M. Rizky Pribadi (2020), yang bertujuan untuk merekam, menghasilkan, dan menginformasikan data sebagai bentuk dukungan terhadap aktivitas pertanian di Indonesia. Penerapan *text mining* agar dapat mengelompokkan data tweet tersebut dengan menggunakan Algoritma *K-Means* yang dalam implementasinya menggunakan 2 tools, yakni orange *tools* untuk *text processing* dan *Rapidminer* untuk melakukan pengolahan algoritma *K-Means*. Dimana ditemukan 5 klaster, yaitu Pangan, Produksi, Lahan, Ekspor, dan Impor. Dimana didapati 2 klaster yang nilainya tinggi yaitu klaster 0 Pangan dengan nilai 0.528% dan klaster 2 Produksi dengan nilai 0.523% sedangakan klaster dengan nilai terendah adalah kluster 3 Ekspor dengan nilai 0.123% yang berarti implementasi text mining dapat dilakukan pada *tools rapidminer*[1].

Penelitian yang dilakukan oleh J Rejito, A Atthariq, dan A S Abdullah (2021) menunjukkan dengan menggunakan metode TF-IDF yang mengubah data tweet dalam bentuk teks menjadi bentuk numerik dapat memudahkan pengolahan data dalam bentuk tabel dan grafik, dimana ditemukan kata yang paling banyak di *retweet* terkait Tokopedia adalah konten terkait kuis berhadiah dan yang paling sedikit adalah konten terkait gaya hidup[8].

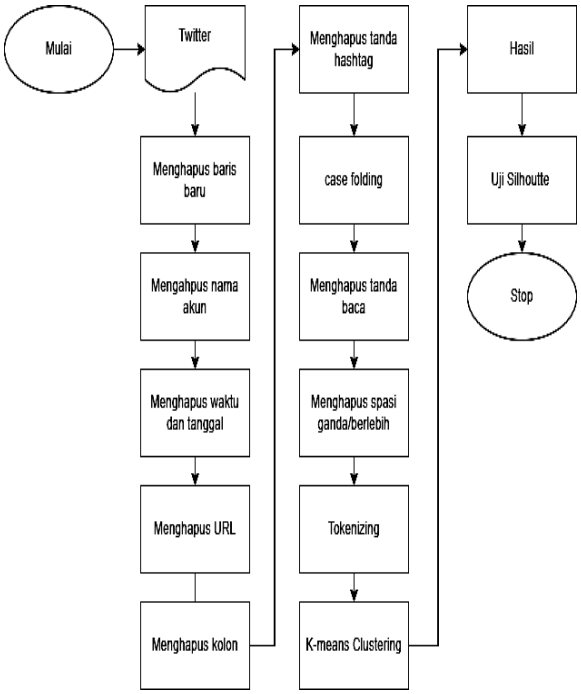
Penelitian yang dilakukan oleh Azizah Nurfauziah Yusril, Inggrit Larasati, dan Qurrotul Aini (2020), yang bertujuan untuk mengetahui jenis konten *tweets* yang banyak diberikan *retweet* dan *favorite* oleh *followers* Gojek Indonesia sehingga dapat digunakan sebagai media advertising kepada pengguna Twitter. Data *tweets* dari Twitter akan dikumpulkan dengan mengintegrasikan Twitter API dan bahasa pemrograman R menggunakan *tools* R Studio. Ditemukan 2 klaster *tweets* setelah dilakukan analisis menggunakan *Text Mining* dan klasterisasi dengan *K-Means.* Dimana diketahui jumlah konten dengan *retweet* terbanyak yaitu terkait program kuis dan pengenalan produk Gojek Indonesia, sehingga Gojek dapat memanfaatkan fitur *retweet*

dan *favorite* sebagai media advertising kepada pengguna Twitter[2].

III. ANALISIS PERMASALAHAN

A. Perancangan sistem

Rancangan dari system clustering pada Health News pada Twitter Data set dengan menggunakan K-Means *Clustering* terdiri dari beberapa tahapan, berikut gambaran yang merupakan beberapa tahapan yang akan dilalui



B. Text Preprocessing

Pada tahap ini dilkaukan beberapa tahap sampai data akan dilakukan clustering. Tahap tahap yang dilakukan adalah sebagai berikut[7]:

a. Menghapus newline/line breaks
   Pada tahap ini berfungi untuk membersihkan data dari line breaks atau newline pada data.
b. Menghapus nama akun
   Nama akun tidak digunakan sehingga bisa dilakukan penghapusan
c. Menghapus tanggal dan waktu
   Tanggal dan waktu tidak digunakan pada penelitian ini tidak digunakan. Maka dilakukan penghapusan.
d. Menghapus URL
e. Menghapus kolon
f. Menghapus tanda pagar atau hashtag
g. Case folding
   case folding adalah perubahan semua huruf bercetak kapital menjadi huruf kecil
h. menghapus punctuation
   penghapusan tanda baca dilakukan karena dalam penelitian ini tanda baca tidak digunakan
i. menghapus white space
   penghapusan white space digunakan untuk mengurangi spasi berlebih.
j. tokenizing
   tokenizing adalah proses merubah tweet menjadi kata per kata(token). Sehingga dapat di lakukan clustering ataupun pembobotan kata

C. K-Means Clustering

Algoritma K-Means bekerja dengan cara membagi data ke dalam k buah cluster yang telah ditentukan[9]. Perhitungan jarak yang digunakan dalam penelitian ini adalah euclidean similarity. Tahap-tahap algoritma dasar K-Means seperti berikut:

1. Menentukan banyaknya k sebagai cluster atau kelompok yang akan dibentuk.
2. Menentukan pusat cluster atau centroid secara acak sebanyak k.
3. Menentukan jarak setiap objek atau dokumen terhadap dokumen lain.

Untuk menghitung jarak tiap dokumen dengan dokumen lain menggunakan Euclidian Distance, dengan persamaan sebagai berikut:

$$d(p,q) = \sqrt{\sum_{t=1}^{nt} (W_{t,p} - W_{t,q})^2}$$

4. Mengelompokkan setiap dokumen berdasarkan kedekatannya dengan centroid(jarak terkecil)
5. Menentukan pusat cluster baru. Memperbaharui nilai centroid dari rata-rata cluster yang bersangkutan. Untuk menentukan centroid baru dapat menggunakan persamaan berikut:

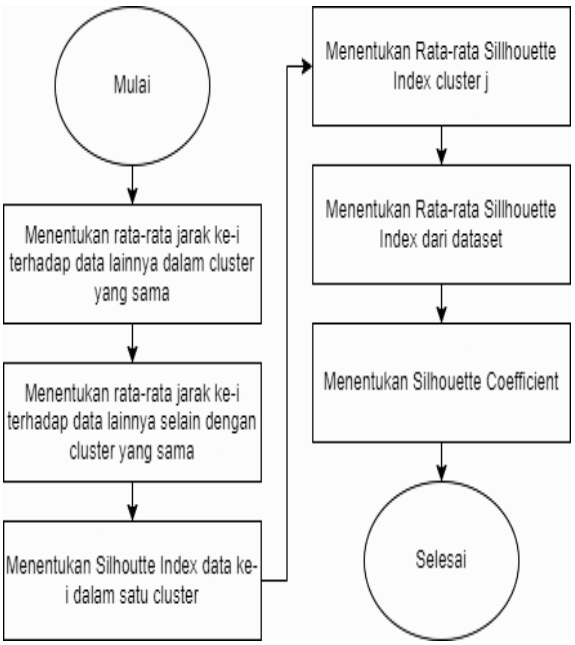
$$C_{k,t} = \frac{\sum_{x=1}^{n} W_{t,x}}{n}$$

6. Mengulangi langkah 3 hingga 5 sampai anggota yang ada pada tiap cluster tidak berubah.
7. Jika langkah 6 sudah terpenuhi, maka centroid pada perulangan terakhir akan digunakan sebagai parameter untuk pengelompokkan dokumen. Kemudian menghitung Silhouette Coefficient dengan k dan anggota-anggota cluster yang didapat.
8. Mengulangi langkah 1 hingga 6 dengan k yang berbeda untuk menghitung Silhouette Coefficient.

D. Pengujian

Silhouette coefficient merupakan salah satu metode evaluasi yang digunakan untuk menguji kualitas dan kekuatan dari sebuah cluster[15].

Berikut merupakan ilustrasi dari diagram alir perhitungan silhouette coefficient



1. Menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen yang berada dalam satu cluster, nilai ini disebut a_t^j.
   $a_t^j = \frac{1}{m_{j-1}} \sum_{\substack{r=1 \\ r \neq 1}}^{n} d\left(x_i^j, x_r^j\right)$
2. Kemudian menghitung rata-rata jarak tiap dokumen ke-i dengan semua dokumen di cluster lain, mengambil nilai terkecil dari semua jarak ratarata tersebut. Nilai ini disebut b_i^j.
   $$b_i^j = min_{\substack{n=1,\dots k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq 1}}^{m_n} d\left(x_i^j, x_r^n\right) \right\}$$
3. Kemudian menghitung nilai silhouette index data ke-i pada satu cluster dengan menggunakan persamaan:
   $SI_i^j = \frac{b_i^j - a_i^j}{\max\left(a_i^j, b_i^j\right)}$
4. Menghitung nilai silhouette index pada cluster j.
   $SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_t^j$
5. Menghitung rata-rata index dari dataset
   $SI = \frac{1}{k} \sum_{j=1}^{k} SI_j$
6. Menentukan silhouette coefficient
   $SC = maks_k SI(k)$

IV. HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan 2 perlakuan yaitu tanpa dan dengan PySpark. Berikut merupakan hasil dari eksperimen tanpa menggunakan library PySpark

Pada eksperimen pertama dengan K = 3, algoritma K-Means konvergen setelah 3 iterasi. Jumlah tweet dalam masing-masing klaster adalah sebagai berikut: klaster 1 memiliki 1394 tweet, klaster 2 memiliki 1559 tweet, dan klaster 3 memiliki 976 tweet. SSE untuk eksperimen ini adalah 3402.023678158529, yang mengukur sejauh mana setiap titik tweet berjarak dari pusat klasternya. Silhouette Score untuk eksperimen ini adalah 0.0027409235213496225, yang mengukur seberapa baik setiap tweet cocok dengan klasternya sendiri dibandingkan dengan klaster tetangga terdekatnya.

```
------ Running K means for experiment no. 1 for k = 3
running iteration 0
running iteration 1
running iteration 2
converged
1:  1394 tweets
2:  1559 tweets
3:  976 tweets
--> SSE : 3402.023678158529
Silhouette Score: 0.0027409235213496225
```

Pada eksperimen kedua dengan K = 4, algoritma K-Means konvergen setelah2 iterasi. Jumlah tweet dalam masing-masing klaster adalah sebagai berikut: klaster 1 memiliki 1416 tweet, klaster 2 memiliki 753 tweet, klaster 3 memiliki 1025 tweet, dan klaster 4 memiliki 735 tweet. SSE untuk eksperimen ini adalah 3345.5032231932355. Silhouette Score untuk eksperimen ini adalah 0.0031310333059951746.

```
------ Running K means for experiment no. 2 for k = 4
running iteration 0
running iteration 1
converged
1:  1416 tweets
2:  753 tweets
3:  1025 tweets
4:  735 tweets
--> SSE : 3345.5032231932355
Silhouette Score: 0.0031310333059951746
```

Pada eksperimen ketiga dengan K = 5, algoritma K-Means konvergen setelah 3 iterasi. Jumlah tweet dalam masing-masing klaster adalah sebagai berikut: klaster 1 memiliki 884 tweet, klaster 2 memiliki 1216 tweet, klaster 3 memiliki 664 tweet, klaster 4 memiliki 470 tweet, dan klaster 5 memiliki 695 tweet. SSE untuk eksperimen ini adalah 3312.6370668736663. Silhouette Score untuk eksperimen ini adalah 0.003479863529015515.

```
------ Running K means for experiment no. 3 for k = 5
running iteration 0
running iteration 1
running iteration 2
converged
1:  884 tweets
2:  1216 tweets
3:  664 tweets
4:  470 tweets
5:  695 tweets
--> SSE : 3312.6370668736663
Silhouette Score: 0.003479863529015515
```

Pada eksperimen keempat dengan K = 6, algoritma K-Means konvergen setelah 5 iterasi. Jumlah tweet dalam masing-masing klaster adalah sebagai berikut: klaster 1 memiliki 1069 tweet, klaster 2 memiliki 490 tweet, klaster 3 memiliki 841 tweet, klaster 4 memiliki 802 tweet, klaster 5 memiliki 437 tweet, dan klaster 6 memiliki 290 tweet. SSE untuk eksperimen ini adalah 3319.101913717965. Silhouette Score untuk eksperimen ini adalah 0.0033334537660398528.

```
------ Running K means for experiment no. 4 for k = 6
running iteration 0
running iteration 1
running iteration 2
running iteration 3
running iteration 4
converged
1:  1069 tweets
2:  490 tweets
3:  841 tweets
4:  802 tweets
5:  437 tweets
6:  290 tweets
--> SSE : 3319.101913717965
Silhouette Score: 0.0033334537660398528
```

Pada eksperimen kelima dengan K = 7, algoritma K-Means konvergen setelah 4 iterasi. Jumlah tweet dalam masing-masing klaster adalah sebagai berikut: klaster 1 memiliki 871 tweet, klaster 2 memiliki 585 tweet, klaster 3 memiliki 732 tweet, klaster 4 memiliki 322 tweet, klaster 5 memiliki 300 tweet, klaster 6 memiliki 673 tweet, dan klaster 7 memiliki 446 tweet. SSE untuk eksperimen ini adalah 3288.3146163817337. Silhouette Score untuk eksperimen ini adalah 0.002839209638317502.

```
------ Running K means for experiment no. 5 for k = 7
running iteration 0
running iteration 1
converged
1:  871 tweets
2:  585 tweets
3:  732 tweets
4:  322 tweets
5:  300 tweets
6:  673 tweets
7:  446 tweets
--> SSE : 3288.3146163817337
Silhouette Score: 0.002839209638317502
```

Sedangkan untuk eksperimen dengan menggunakan pyspark emiliki hasil sebagai berikut:

```
------ Running K-means for experiment no. 1 for k = 3
Running iteration 0
Converged
--> SSE: 6000


------ Running K-means for experiment no. 2 for k = 4
Running iteration 0
Converged
--> SSE: 11611


------ Running K-means for experiment no. 3 for k = 5
Running iteration 0
Converged
--> SSE: 24922


------ Running K-means for experiment no. 4 for k = 6
Running iteration 0
Converged
--> SSE: 39252


------ Running K-means for experiment no. 5 for k = 7
Running iteration 0
Converged
--> SSE: 54002
```

Pada penerapan library PySpark ternyata hanya melakukan 1 kali iterasi dan pada iterasi ke-2 sudah dalam keadaan konvergen. Untuk k=3 SSE bernilai 6000, k=4 SSE=11611, k=5 SSE=24922, k=6 SSE=39252 dan k=7 SSE=54002. Dalam masing-masing eksperimen nilai Sum of squared error (SSE) bertambah dengan sangat signifikan yang artinya bahwa kualitas clustering yang dilakukan masih tergolong rendah.

SSE digunakan untuk mengukur sejauh mana setiap titik tweet berjarak dari pusat klasternya. Semakin rendah SSE, semakin baik kualitas klasteringnya. Dalam eksperimen ini, SSE cenderung berkurang saat nilai K meningkat.

Silhouette Score digunakan untuk mengukur seberapa baik setiap tweet cocok dengan klasternya sendiri dibandingkan dengan klaster tetangga

terdekatnya. Nilai Silhouette Score yang rendah menunjukkan bahwa klasterisasi tidak begitu baik. Dalam eksperimen ini, nilai Silhouette Score cenderung rendah dan hampir sama di setiap eksperimen, menunjukkan bahwa klasterisasi mungkin tidak optimal.

Berdasarkan hasil eksperimen, tidak ada nilai K yang memberikan klasterisasi yang sangat baik. Mungkin perlu eksperimen lebih lanjut dengan nilai K yang berbeda atau menggunakan algoritma klasterisasi yang berbeda untuk mendapatkan hasil yang lebih baik. Selain itu, penting untuk menganalisis ciri-ciri dan konteks data tweet yang digunakan untuk memahami mengapa hasil klasterisasi tidak optimal.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian yang sudah didapatkan, dapat disimpulkan bahwa penerapan Pyspark pada dataset yang dimiliki tidak menunjukan hasil yang lebih baik dikarenakan Nilai evaluasi SSE (Sum of Squared Error) yang bernilai 3288.3146163817337 hingga 3402.023678158529.

B. Saran

Berdasarkan penelitian yang telah dilakukan, berikut beberapa saran perbaikan ataupun pengembangan yang dapat dilakukan pada penelitian kedepannya :

1. Eksplorasi nilai K yang berbeda, Praproses data yang lebih cermat, seperti menghilangkan kata-kata yang tidak relevan, memperbaiki kesalahan pengejaan, atau mengurangi dimensi data, untuk memperoleh representasi yang lebih baik. Menggunakan metode klasterisasi yang berbeda yang mungkin memberikan hasil yang lebih baik terhadap data tweet tersebut.

REFERENSI

[1] H. Irsyad and M. R. Pribadi, "Implementation of Text Mining in Clustering Indonesian Agricultural Tweets Data Using K-Means," in KURAWAL Jurnal Teknologi, Informasi dan Industri, vol. 3, no. 2, 2020.

[2] A. N. Yusril, I. Larasati, and Q. Aini, "Text Mining Implementation for Advertising Using K-Means Clustering Method on Gojek Indonesia Tweets Data," in SISTEMASI: Jurnal Sistem Informasi, vol. 9, no. 3, 2020.

[3] D. A. C. Rachman, R. Goejntoro, and F. D. T. Amijaya, "Implementation of Text Mining in Clustering Thesis Documents Using K-Means Clustering Method," in Jurnal EKSPONESIAL, vol. 11, no. 2, 2020.

[4] Y. Darmi and A. Setiawan, "Application of K-Means Clustering Method in Product Sales Clustering," in Jurnal Media Infotama, vol. 12, no. 2,
2016.

[5] N. Gherabi and J. Kacprzyk, "Intelligent System in Big Data, Semantic Web, and Machine Learning," in Proceedings of the 2023 International Conference on Intelligent Systems (ICIS), vol. 1344.

[6] G. E. I. Kambey, R. Sengkey, and A. Jacobus, "Clustering Implementation in Indonesian Text Document Similarity Detection Application," in Jurnal Teknin Informatika, vol. 15, no. 2, 2020.

[7] M. Alsolamy, A. Alabbas, A. M. Alotaibi, and M. Abdullah, "TopicBased Sentiment Analysis for COVID-19 Tweets," in International Journal of Advanced Computer Science and Applications, 2021.

[8] J. Rejito, A. Atthariq, and A. S. Abdullah, "Application of Text Mining Employing K-Means Algorithms for Clustering Tokopedia Tweets," in Journal of Physics: Conference Series, 2021.

[9] J. Rashid, S. M. A. Shah, A. Irtaza, et al., "Topic Modeling Technique for Text Mining over Biomedical Text Corpora through Hybrid Inverse Document Frequency and Fuzzy K-Means Clustering," in Proceedings 39 of the 2019 International Conference on Data Science and Advanced Analytics (DSAA), pp. 123-128.

[10] N. S. Dhuha, "Text Classification of Online Complaints Using Support Vector Machine (SVM)," in Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 4, 2020.

[11] V. K. Bakti and A. Rakhman, "Clustering of Research Documents in Higher Education Using K-Means Clustering for System Implementation of Information Retrieval," in Smart Comp, vol. 10, no. 3, 2021.

[12] A. Ryansyah and S. Andayani, "Impelementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen," Jurnal Sistem & Teknologi Informasi Komunikasi, vol. 1, no. 1.

[13] B. M. Metisen and H. L. Sari, "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila," Jurnal Media Infortama, vol. 11, no. 2, 2015.

[14] J.-H. Yu and Z.-M. Zhou, "Components and

Development in Big Data System: A Survey," Journal of Electronic Science and Technology, vol. 17, no. 1, 2019.
[15] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coeficient Untuk Evaluasi Clustering Obat," PENA TEKNIK Jurnal Ilmiah – Ilmu Teknik, vol. 6, no. 2, 2021