Journal of
Information Systems Engineering
and Business Intelligence

Vol.X, No.X, February/June/October XXXX
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Klasterisasi dan Geovisualisasi *Tweet* Penyebaran Penyakit Menular Langsung: Studi Kasus COVID-19

Fahmirullah Abdillah [1)*], Ira Puspitasari [2)], Eto Wuryanto[3)]

[1)2)3)] *Information Systems, Faculty of Science and Technology, Universitas Airlangga, IndonesiaJl. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya*

[1)] fahmirullah.abdillah-2018@fst.unair.ac.id, [2)]ira-p@fst.unair.ac.id, [3)] eto-w@fst.unair.ac.id

Abstract

Background: Layanan media sosial *microblog* seperti Twitter menghasilkan aliran besar dalam penyebaran informasi terhadap suatu kejadian. Media sosial dimanfaatkan sebagai bentuk pengawasan penyebaran penyakit menular langsung.
Objective: Penelitian ini bertujuan untuk menerapkan algoritma *clustering* berbasis kepadatan, algoritma *Density-based Spatial Clustering Applications with Noise* (DBSCAN) dan *Ordering Points to Identify the Clustering Structure* (OPTICS).
Methods: Klasterisasi dilakukan dengan memasukkan data hasil *scraping, preprocessing*, dan pembobotan TF-IDF. Klasterisasi dilakukan pada data berjumlah 8.114 sampel dan menggunakan parameter *epsilon, xi score*, dan *minpts*. Proses uji validasi klaster menggunakan *silhouette coefficient*.
Results: Klasterisasi terbaik diperoleh dari algoritma OPTICS yang menghasilkan 6 klaster, dengan nilai *silhouette coefficient* sebesar 0,6508317895 dan setiap klaster memiliki *term* terbaik.
Conclusion: Metode klasterisasi menggunakan algoritma DBSCAN dan OPTICS dalam mengelompokkan gejala penyakit menular langsung dengan pendekatan berbasis kepadatan terbukti cukup baik. Penilaian ini dibuktikan pada hasil uji validasi klaster yang baik pada algoritma OPTICS.

Keywords: Twitter, *Preprocessing*, *Clustering*, DBSCAN, OPTICS, *Density-based Algorithm, silhouette coefficient*.

Article history: Received 5 April 20XX, first decision 22 April 20XX, accepted 22 August 20XX, available online 28 October 20XX

I. INTRODUCTION

Layanan media sosial mikroblog seperti Twitter menghasilkan aliran besar dalam penyebaran informasi terhadap suatu kejadian. Sumber informasi realtime ini sangat berharga untuk banyak area aplikasi, khususnya untuk deteksi bencana dan skenario respons. Terbukti dengan aliran volume maupun kecepatan tweet saat kejadian berlangsung sangat tinggi dan cepat, sehingga masyarakat yang terdampak maupun petugas profesional sedikit mengalami kesulitan saat pemrosesan informasi [11]. Melalui pemantauan tweet dapat dideteksi adanya gempa bumi. Probabilitas yang dihasilkan oleh Japan Meteorology Agency cukup tinggi, yaitu 96% untuk gempa bumi dengan skala richter 3 atau lebih [19], maka dari itu situs mikroblog ini dapat digunakan sebagai sistem sensor untuk mendeteksi suatu bencana alam atau kejadian lainnya [4].

Beberapa penelitian yang menggunakan data dari media sosial Twitter telah dilakukan sebelumnya. Salah satunya, melakukan penelitian tentang akuisisi dan klasterisasi data teks Twitter untuk memperoleh dasar pengetahuan terhadap profil pengguna Twitter. Penelitian dilakukan dengan ujicoba keyword “K-Pop” dan “K-Drama”. Dari hasil ujicoba akuisisi data didapatkan sebanyak 68.393 tweet. Hasil tersebut disebar menjadi 3 klaster / k=3, yang mana klaster pertama adalah waktu tweet dianggap pada pagi hari, klaster kedua adalah waktu tweet dianggap pada siang hari, dan klaster ketiga adalah waktu tweet dianggap pada malam hari. Kemudian, hasil klasterisasi didapat jam 21.00 - 01.00 merupakan mayoritas orang-orang melakukan tweet. Dari hasil penelitian ini kita dapatkan bahwa penentuan nilai k untuk memperkirakan topik suatu klaster didasarkan pada asumsi kebiasaan pengguna dalam menggunakan media sosial Twitter [6].

Penelitian lainnya tentang kemungkinan analisis secara realtime pada media sosial dan otomatis dari pesan Twitter selama terjadinya situasi darurat [25]. Analisis dilakukan menggunakan tool ekstraksi informasi yang berhasil mendapatkan 97.000 tweet yang dikirim sebelum, saat, dan setelah kejadian alam (badai) terjadi. Lokasi kejadian adalah di Belgia saat berlangsungnya festival Pukkelpop di tahun 2011. Tool ekstraksi dapat menganalisis tweet melalui tampilan geografis, jenis isi pesan (kerusakan, korban), dan jenis tweet (seperti retweet).

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.X.X.pp-pp

Penyakit menular langsung merupakan suatu infeksi yang disebabkan oleh mikroorganisme, seperti virus, parasit, atau jamur. Infeksi ini dapat berpindah dari orang yang sakit ke orang yang sehat. Bentuk penularannya bisa terjadi secara langsung maupun tidak langsung, penularan secara langsung terjadi ketika benda tak kasat mata di atas pada orang yang sakit berpindah melalui kontak fisik, misalnya lewat sentuhan. Saat ini penyakit menular langsung telah menjadi wabah yakni virus Covid-19. Wabah yang terjadi secara mendunia ini diberi nama Coronavirus Disease 2019 (Covid-19) yang disebabkan oleh Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). Penyebaran penyakit menular langsung ini hingga ke seluruh penjuru nusantara dan dunia. Virus ini dapat ditularkan dari manusia ke manusia dan telah menyebar secara luas di China (sebagai tempat kemunculan pertama) dan lebih dari 190 negara dan teritori lainnya. Pada 12 Maret 2020, WHO mengumumkan COVID-19 sebagai pandemi. Hingga tanggal 29 Maret 2020, terdapat 634.835 kasus dan 33.106 jumlah kematian di seluruh dunia. Sementara di Indonesia sudah ditetapkan 1.528 kasus dengan positif COVID-19 dan 136 kasus kematian. Per tanggal 20 Desember 2020, Satgas Covid-19 menerbitkan laporan yang berisi informasi kasus terkonfirmasi positif, sembuh, ataupun meninggal. Sebanyak 735.124 kasus terkonfirmasi positif dan 19.880 (2,99%) jumlah kematian di Indonesia, serta jumlah kasus sembuh 541.811 (81,48%) [23].

Metode Density-Based Spatial Cluster of Application with Noise (DBSCAN) merupakan salah satu metode cluster mengacu pada densitas atau kepadatan. Kepadatan yang dimaksudkan yaitu dalam metode DBSCAN mengelompokkan wilayah dengan jarak yang telah ditentukan menggunakan nilai parameter Epsilon dan MinPts, sehingga dihasilkan suatu kelompok yang padat dengan jarak antar anggota kelompok yang beragam. Parameter Epsilon merupakan jarak maksimal antar titik pusat dengan titik anggota dalam suatu cluster. Sedangkan MinPts merupakan minimal anggota yang harus terpenuhi dalam sebuah klaster. Apabila kedua parameter tersebut telah terpenuhi, maka akan terbentuklah suatu klaster [5]. Analisis *cluster* merupakan teknik multivariat dalam analisis statistik yang dapat mengumpulkan objek-objek dengan karakteristik sama pada suatu kelompok yang lebih kecil. Pada penelitian ini metode klasterisasi *tweet* yang digunakan adalah algoritma *Density-based Spatial Clustering Applications with Noise* (DBSCAN) dan *Ordering Points to Identifying the Clustering Structure* (OPTICS). Metode-metode ini dipilih dan dibandingkan karena keduanya dapat menghasilkan *cluster* tanpa penentuan *centroids* dan juga dapat menemukan titik-titik yang menyimpang. Dataset yang digunakan berisi 8.114 sampel data yang diperoleh dari 15 kata kunci. Data hasil klasterisasi divisualisasikan untuk menerapkan geovisualisasi *tweet* untuk kasus penyebaran penyakit menular langsung (studi kasus Covid-19). Proses geovisualisasi digunakan untuk mendapatkan hasil tampilan data tweet hasil klasterisasi dan lokasi penyebaran *tweet* terkait penyebaran penyakit menular langsung (studi kasus Covid-19). Pengujian dilakukan dengan mengevaluasi hasil analisis klasterisasi menggunakan nilai *silhouette coefficient*.

II. Literature Review

Pada penelitian sebelumnya, media Twitter sering digunakan untuk meneliti terhadap suatu kejadian. Informasi yang terkandung pada media Twitter dapat digunakan sebagai rekomendasi kebijakan yang cukup untuk meningkatkan kesadaran masyarakat [6][23][24].

Crooks et al. [4] melakukan analisis performa *microblogging* sebagai sistem sensor untuk mendeteksi kejadian dengan studi kasus gempa bumi yang ada di daerah East Coast, Amerika Serikat. Peneliti mengambil hasil deteksi yang memiliki karakteristik spasial dan temporal dari penyebaran informasi yang ada di situs *microblogging* (Twitter). Analisis terhadap situs ini juga dilakukan dengan teknik *crowdsourcing*, karena setiap media sosial atau situs microblogging juga memiliki informasi geografis ketika seorang pengguna mengomentari suatu kejadian yang dialami terjadi di sekitarnya, atau mengenai pusat lokasi yang menjadi pusat perhatian. Namun, perbedaannya media sosial atau situs *microblogging* tidak menyediakan informasi geografis pengguna secara terang-terangan, berbeda dengan teknik crowdsourcing yang sudah ada pada aplikasi Wikimapia atau OpenStreetMap. Penelitian ini bertujuan untuk menilai kualitas informasi yang telah diambil dari masyarakat dengan mempertimbangkan reaksi pengguna Twitter terhadap gempa bumi yang terjadi di Virginia, Amerika Serikat pada tanggal 23 Agustus 2011. Hasilnya, *tweet* dapat digunakan untuk memberi perkiraan yang cepat dan bagus dari wilayah yang terkena dampak gempa bumi. Perkiraan ini digunakan sebagai informasi yang penting untuk penanganan dan pemulihan dampak bencana. Dengan kemampuannya untuk memperkirakan wilayah yang terkena dampak gempa bumi dengan akurat, hal tersebut mendukung pernyataan bahwa dengan mengambil informasi geospasial di Twitter, peneliti memperoleh informasi yang penting mengenai dampak dari suatu kejadian dengan cepat.

Wahyuni et al. [25] melakukan perhitungan pembobotan dalam frekuensi kemunculan sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata yang ingin diteliti. Frekuensi kemunculan kata yang menunjukkan seberapa penting kata tersebut dalam kumpulan dokumen. Rumus ini disebut dengan algoritme TF-IDF.

Penelitian selanjutnya, mengidentifikasi informasi tentang resiko dan respon komunikasi masyarakat Indonesia terhadap pemberlakuan *New Normal* ketika pandemi Covid-19 yang ada pada situs *microblogging* (Twitter) di wilayah Indonesia. Penelitian ini bertujuan untuk menggolongkan *tweet* yang memiliki sentimen positif, negatif, dan netral dengan klasifikator *naïve-bayes* dan memasukkannya ke dalam analisis emosi dasar dari *Plutchik's Wheel of Emotions* (*joy, fear, anticipation, anger, disgust, sadness, surprise,* dan *trust*). Penelitian ini dilakukan pada tanggal 21 Mei 2020 – 18 Juni 2020, dengan hasil data sebanyak 282.216 *tweet* dari 137.057 pengguna. *Tweet* itu semua mengandung 88.677 *mention*, 31.452 *reply*, 164.087 *retweet*. Hasil tersebut disebar ke dalam *Plutchik's Wheel of Emotions* dengan persentase; *joy* (9,01%), *fear* (6,50%), *anticipation* (14,82%), *anger* (4,81%), *disgust* (0,73%), *sadness* (1,74%), *surprise* (8,62%), dan *trust* (53,77%). Kemudian, didapatkan hasil penggunaan tiga *hashtag* terbanyak, yaitu *#NewNormal* (17.051 *tweet*), *#TataKehidupanBaru* (10.980 *tweet*), dan *#DisiplinPolaHidupBaru* (5.200 *tweet*). Dari hasil ini [18] peneliti dapat menggolongkan suatu kejadian yang ada di situs *microblogging* (Twitter) ke dalam pemetaan analisis berdasarkan emosi pengguna.

Kajian-kajian tersebut di atas jelas mempunyai fokus dan domain tersendiri. Semuanya menyentuh beberapa poin dalam klasterisasi dengan DBSCAN dan OPTICS yang termasuk dalam *density-based clustering*. Semuanya secara implisit menunjukkan keterkaitan keduanya. Namun, belum ada satupun yang mengeksplorasi performansi validasi antara klasterisasi dengan DBSCAN dan OPTICS.

III. METHODS

Alur yang digunakan pada penelitian ini dimulai dengan tahap melakukan pengumpulan data, kemudian dilanjutkan dengan data tersebut diproses yang dapat dilihat pada diagram Gambar 1.

Gambar 1

Tahapan tersebut adalah sebagai berikut:

A. Pengumpulan Data
B. *Preprocessing Data*
C. Pembobotan TF-IDF
D. DBSCAN *Clustering*
E. OPTICS *Clustering*
F. Geovisualisasi
G. Uji Validasi / Evaluasi

The methods sections often come disguised with other article-specific section titles, but serve a unified purpose: to detail the methods used in an objective manner without introduction of interpretation or opinion. The methods sections should tell the reader clearly how the results were obtained. In addition, the procedure must be written chronologically and clearly. They should be specific. They should also make adequate reference to accepted methods and identify differences. In the method section, authors are recommended to cite a source who helped in the selection of the method. Authors are expected to describe how results will be measured, tested, and evaluated.

A. Figure

All figures should be numbered with Arabic numerals (1,2,3,….). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed as close as possible to the first reference to them in the paper. The allowed figure forms are picture, vector, and graphic.

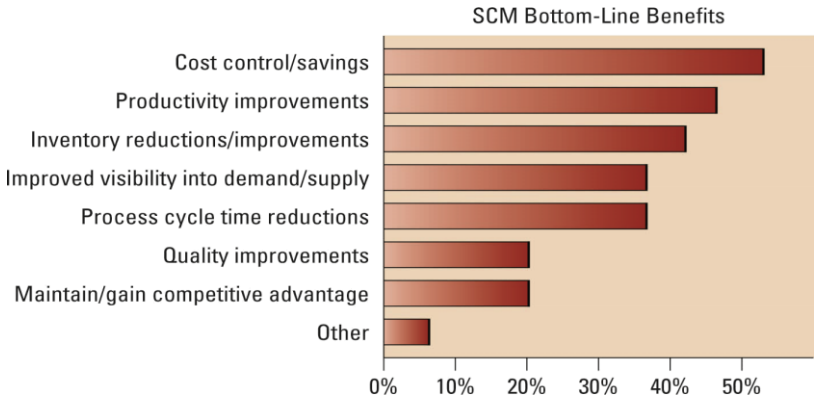


Fig. 1 Figure captions should be standalone. Descriptive enough to be understood without having to refer to the main text.

1) Picture

Preferred format of figures/images are PNG, and JPEG. Please ensure that all the figures are of 330 PPI resolutions as this will facilitate good output. Please ensure that the resolutions in images in word documents are at least in HD (330 ppi: good quality for high-definition displays). Please click on photos and choose "Picture Format" → "Compress Pictures" → "Resolution options" to facilitate this requirement.

The figure number and caption should be typed below the illustration in 8 pt and left justified [Note: one-line captions of length less than column width (or full typesetting width or oblong) centered]. Artwork has no text along the side of it in the main body of the text. For example, see Fig. 1.

2) Vector

Suppose the figures come from a vector drawing app, for example, Visio, Inkscape, Diagrams.net, or other similar applications. In that case, it is better if the figure files in vector format are embedded directly into the manuscript without being converted to pictures. So, the quality of the figure resolution can be maintained. Usually, this format is suitable for describing research procedures in the methods section. For example, see Fig. 2.

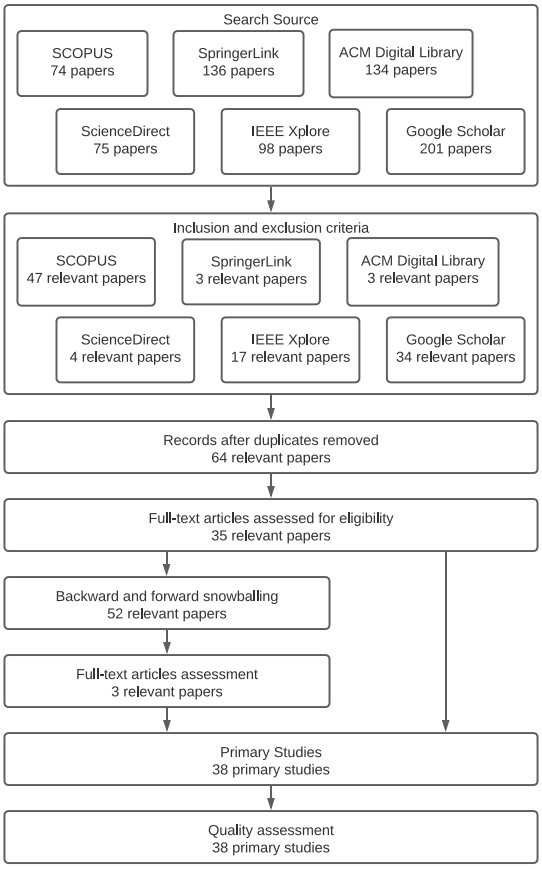
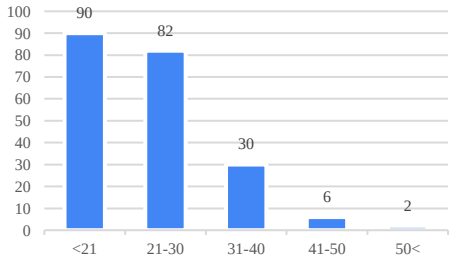


Fig. 2 Study search and selection process.

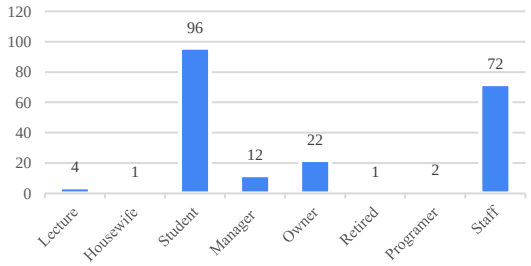
3) Graphic

The same happens if the figures are a graphic made using Microsoft Office. It is perfect if you embed the graph directly into Word so that the resolution quality of the figure remains at its best. Make sure all labels are visible in the document. Please remove the outline lines for the best appearance.

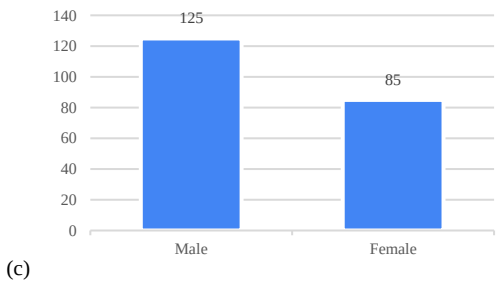
If figures have the same context, they should be placed next to each other and grouped into one caption. For example, see Fig. 3.



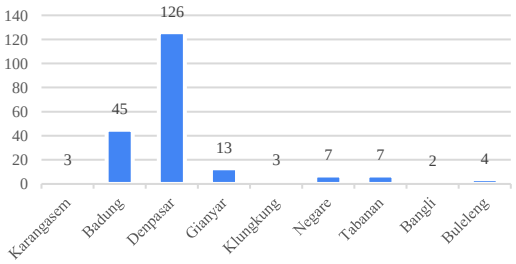
(a)



(b)



(c)



(d)

Fig. 3 Respondent profiles differentiated based on (a) Age; (b) Occupation; (c) Gender; (d) District

B. Table

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Table 1 is an example which the authors may find useful.

TABLE 1
THE SIGNIFICANCE OF THE RELATIONSHIPS IN THE MODEL

Relationships	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P Values*
Perceive Ease of Use ->Attitude	0.286	0.288	0.058	4.907	0.000
Information Quality ->Intention to Use	0.175	0.174	0.072	2.434	0.015
Intention to Use ->Use	0.657	0.658	0.039	17.054	0.000
Use ->Net Benefits	0.463	0.461	0.061	7.576	0.000
Use ->User Satisfaction	0.405	0.400	0.068	5.936	0.000
Performance Expectancy -> Intention to Use	0.052	0.053	0.058	0.893	0.372
User Satisfaction ->Net Benefits	0.428	0.430	0.059	7.207	0.000

*alpha=0.05 (this is additional legend/caption for clarity of data description, if needed)

C. Algorithm

Pseudocode, or structured English, allows a programmer to use English-like sentences to write an explanation of what a program is supposed to do. An example of pseudocode can be seen in Algorithm 1.

Algorithm 1
Person identification

```
function person_identification ()
  get subject and object token
  for each subject and object token
    set token as person
  end for
  return person
```

D. Equations

Equations and formulae should be typed in Mathtype or any Equation Editor, and numbered consecutively with Arabic numerals in parentheses on the right hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space.

$$\alpha + \beta = \chi. \quad (1)$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ..."

IV. RESULTS

A. Hasil pengumpulan
B. Hasil preprocessing data
C. Hasil TF-IDF
D. Hasil DBSCAN Clustering
E. Hasil OPTICS Clustering
F. Hasil geovisualisasi
G. Hasil uji validasi
H. Hasil wordcloud klaster terbaik

The results section and the following discussion section allow the most flexibility in terms of organization and content. In general, the pure, unbiased results should be presented first without interpretation. These results should present the data or the results after applying the techniques outlined in the methods section. The results are simply results; they do not draw conclusions. In the results section, the author must write down the research results in logical sequences, according to the research flow. The study results are presented in the form of narrative/textual, tables, or images in the form of graphs or diagrams. Avoid displaying raw data. The authors are required to clearly describe the evaluation result of the study in this section.

The main purpose of the results section is to provide the data from the study so that other researchers can draw their own conclusions and understand fully the basis for the conclusions. A common format for the results section is to present a series of figures and to describe the figures in detail through the text. A good results section presents clear figures with efficient text. The figures should support the assertions in the paper or illustrate the new insights. Where applicable, results should be illustrated in terms of non-dimensional variables.

V. DISCUSSION

Tujuan dari penelitian ini adalah menghasilkan performa terbaik antara algoritma *clustering* DBSCAN dan OPTICS yang digunakan pada data pada data *tweet* tentang penyebaran penyakit menular langsung dengan studi kasus pandemi COVID-19. Klasterisasi menggunakan algoritma DBSCAN menggunakan beberapa sampel yang menghasilkan parameter optimal pada eps = 1,35 dan minpts = 10. Kemudian, pada algoritma OPTICS menghasilkan parameter optimal pada pada *xi score* = 0,05 dan minpts = 10.

Hasil uji validasi klaster terbaik diperoleh algoritma OPTICS dengan menghasilkan nilai *silhouette coefficient* sebesar 0,6508317895 terbentuk 6 klaster, 1655 *noise*. Sedangkan algoritma DBSCAN menghasilkan nilai *silhouette coefficient* sebesar 0,0056951958 terbentuk 1 klaster, 19 *noise*.

The discussion section is where the article interprets the results to reach its major conclusions. This is also where the author's opinion enters the picture. The discussion is where the argument is made. Common features of the discussion section include: (1) Provide an accurate assessment of the insights gained from the study. (2) The author needs to compare results with other studies. (3) Effectively contextualize the results within the existing body of knowledge. (4) Acknowledge any limitations or constraints encountered during the research (limitations of study or threats to validity). (5) Offer specific and well-justified recommendations for fellow researchers.

It is important to avoid the following pitfalls: (1) Repetition of results already presented. (2) Introduction of new results that were not previously described in the methods or results sections. (3) Introducing relevant literature that should have been discussed earlier. (4) Overemphasizing the significance of findings, especially in cases where statistical significance was not achieved. (5) Downplaying or neglecting the significance of findings when valuable lessons can be derived. (6) Making overly ambitious or overly cautious suggestions for future research or applications.

VI. CONCLUSIONS

Dari penelitian ini dapat disimpulkan bahwa algoritma *clustering* DBSCAN dan OPTICS dapat dilakukan dengan baik, dan algoritma OPTICS menghasilkan performa yang baik dalam proses *clustering*. Hal ini ditunjukkan dengan nilai *silhouette coefficient* yang lebih besar dan mendekati 1.

Author Contributions: *Abdillah*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing. *Puspitasari*: Conceptualization, Supervision, Investigation, Data Curation. *Wuryanto*: Conceptualization, Methodology, Supervision, Investigation, Data Curation.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: *Puspitasari* is the member of Editorial Boards, but had no role in the decision to publish this article. No other potential conflict of interest relevant to this article was reported.

Data Availability: The source code of the article is open access at *Abdillah's* GitHub profile.

Informed Consent: There were no human subjects in this article.

Animal Subjects: There were no animal subjects in this article.

ORCID:
First Author: -
Second Author: https://orcid.org/0000-0001-5983-6257
Third Author: https://orcid.org/0000-0001-5871-0425

REFERENCES

We suggest there should be at least 20 references within the manuscript. Make sure you use Mendeley feature in Microsoft Word for handling citation in manuscript. Please use IEEE Journal standard for Reference style in Mendeley. References may not include all information; please obtain and include relevant information. Title of paper with only first word capitalized. Do not combine references. There must be only one reference with each number. Please include DOI information if applicable.

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ...".

[1] Baumgartner, C., & Graber, A. (2007). Data mining and knowledge discovery in metabolomics. *Successes and New Directions in Data Mining*, *39*(11), 141–166. https://doi.org/10.4018/978-1-59904-645-7.ch007

[2] Budiman, S., Safitri, D., & Ispriyanti, D. (2016). Perbandingan Metode K-Means Dan Metode Dbscan Pada Pengelompokan Rumah Kost Mahasiswa Di Kelurahan Tembalang Semarang. *Jurnal Gaussian*, *5*(4), 757–762.

[3] Chakrabarti, S., Ester, M., Fayyad, U., & Gehrke, J. (2006). Data mining curriculum: a proposal. *Acm Sigkdd*, 1–10. http://pdf.aminer.org/000/303/279/decision_tree_construction_from_multidimensional_structured_data.pdf%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+mining+curriculum:+A+proposal+(Version+1.0)#4%5Cnhttp://scholar.google.com/scholar

[4] Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

[5] Devi, A. S., Putra, I. K. G. D., & Sukarsa, I. M. (2015). Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *6*(3), 185. https://doi.org/10.24843/lkjiti.2015.v06.i03.p05

[6] Dwiarni, B. A., & Setiyono, B. (2019). Akuisisi dan Clustering Data Sosial Media Menggunakan Algoritma K-Means sebagai Dasar untuk Mengetahui Profil Pengguna. *Jurnal Sains Dan Seni*, *8*(2), 2337–3520. https://apps.twitter.com/

[7] Fay, D. L. (1967). 済無No Title No Title No Title. Angewandte Chemie International Edition, 6(11), 951–952.

[8] Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. https://doi.org/10.18637/jss.v025.i05

[9] Freeman, J. (2019). What is an API? Application programming interfaces explained. In *InfoWorld* (pp. 1–9). https://www.infoworld.com/article/3269878/what-is-an-api-application-programming-interfaces-explained.html

[10] Han, J., Kamber, M., & Pei, J. (Eds.). (2012). About the Authors. In *Data Mining (Third Edition)* (Third Edit, p. xxxv). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00027-7

[11] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster- Related messages in social media. *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management*, *May*, 791–801.

[12] Koko Mukti Wibowo, Indra Kanedi, J. J. (2021). Sistem Informasi Geografis (Sig) Menentukan Lokasi Pertambangan Batu Bara Di Provinsi Bengkulu Berbasis Website. *Jurnal Media Infotama*, *11*(1), 223–260.

[13] Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311. https://doi.org/10.1016/j.eswa.2012.02.063

[14] Melcer, E. F., & Isbister, K. (2018). Bots & (main)frames: Exploring the impact of tangible blocks and collaborative play in an educational programming game. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*(April). https://doi.org/10.1145/3173574.3173840

[15] Nurdiana, O., Jumadi, J., & Nursantika, D. (2016). Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia. *Jurnal Online Informatika*, *1*(1), 59. https://doi.org/10.15575/join.v1i1.12

[16] Prabahari, R. . T. (2014). A Comparative Analysis of Density Based Clustering Techniques for Outlier Mining. 3(11), 132–136.

[17] Putri, M. M., Dewi, C., Permata Siam, E., Asri Wijayanti, G., Aulia, N., & Nooraeni, R. (2021). *Komparasi DBSCAN dan K-Means Clustering pada Pengelompokan Status Desa di Jawa Tengah Tahun 2020*. *17*(3), 394–404. https://doi.org/10.20956/j.v17i3.11704

[18] Rahmanti, A. R., Ningrum, D. N. A., Lazuardi, L., Yang, H. C., & Li, Y. C. (2021). Social Media Data Analytics for Outbreak Risk Communication: Public Attention on the "New Normal" During the COVID-19 Pandemic in Indonesia. *Computer Methods and Programs in Biomedicine*, *205*, 106083. https://doi.org/10.1016/j.cmpb.2021.106083

[19] Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 919–931. https://doi.org/10.1109/TKDE.2012.29

[20] Salman, N. (2023). Density-Based Clustering Analysis. 8, 1–8.

[21] Santoso, A. M. . (2022). Covid-19 : Varian Dan Mutasi. *Jurnal Medika Hutama*, *3*(02), 1980–1986. https://jurnalmedikahutama.com/index.php/JMH/article/view/396/271

[22] Silitonga, P. (2016). ANALISIS POLA PENYEBARAN PENYAKIT PASIEN PENGGUNA BADAN PENYELENGGARA JAMINAN SOSIAL (BPJS) KESEHATAN DENGAN MENGGUNAKAN METODE DBSCAN CLUSTERING (Studi Kasus Rumah Sakit Umum Pusat Haji Adam Malik Medan). *Jurnal TIMES*, *Vol. V No*(ISSN : 2337-3601), 11–40. http://etd.lib.metu.edu.tr/upload/12620012/index.pdf

[23] Susanto, H., Sumpeno, S., & Rachmadi, R. F. (2014). Visualisasi Data Teks TwitterBerbasis Bahasa Indonesia Menggunakan Teknik Pengklasteran. *Jurnal Teknik Elektro Institut Teknologi Sepuluh Nopember*, 6. http://digilib.its.ac.id/ITS-paper-22121150006831/35629

[24] Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yunihastuti, E. (2020). Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, *7*(1), 45. https://doi.org/10.7454/jpdi.v7i1.415

[25] Wahyuni, R. T., Prastiyanto, D., & Supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro Universitas Negeri Semarang*, *9*(1), 18–23. https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659.

[26] Terpstra, Teun & Vries, A. & Stronkman, Richard & Paradies, G.L.. (2012). Towards a realtime twitter analysis during crises for operational crisis management. 1-9.