Ekstraksi Fitur Produktivitas Dinamis berdasarkan Topik Artikel Ilmiah untuk Klasterisasi Peneliti

Extraction of Productivity Features based on Scientific Article Topics for Research Clusterization

Addien Haniefardy[1], Diana Purwitasari[2], Chastine Fatichah[3]
[1,2,3] Departemen Teknik Informatika, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember (FTEIC - ITS)
E-mail: [1]haniefardy12@mhs.if.its.ac.id,[2]diana@if.its.ac.id, [3]chastine@if.its.ac.id

Abstrak

Pengelompokkan peneliti seringkali menggunakan informasi tekstual yang terdapat pada artikel ilmiah peneliti, contohnya judul, abstrak, dan kata kunci sehingga menghasilkan kelompok peneliti dengan kemiripan informasi tekstual pada artikel ilmiah mereka. Pengelompokkan peneliti juga seringkali menggunakan jumlah publikasi dan sitasi sehingga menghasilkan kelompok peneliti yang memiliki jumlah publikasi dan sitasi yang cenderung sama. Berdasarkaan kedua metode di atas, penelitian ini mencoba untuk menganalisis penggunaan topik artikel ilmiah pada proses ekstraksi fitur produktivitas. Fitur ini merupakan fitur yang didapatkan melalui penghitungan kinerja peneliti berdasarkan jumlah publikasi dan sitasi. Hasil ekstraksi fitur akan digunakan untuk klasterisasi peneliti menggunakan metode K-Means++. Sebelum data peneliti diklasterisasi, terlebih dahulu data peneliti dianalisis untuk menghilangkan kemungkinan adanya *outlier*. Evaluasi hasil klaster dilakukan dengan mempertimbangkan nilai *sum squared error* dan s*ilhouette*. Hasilnya, klaster optimal didapatkan dengan nilai K sama dengan 8 dan nilai *silhouette* sama dengan 0.15396. Kemudian, hasil klaster dianalisis untuk dapat memberikan label terhadap masing-masing klaster dengan mempertimbangkan topik artikel ilmiah, jumlah publikasi dan jumlah sitasi.

Kata kunci: Klasterisasi peneliti, Produktivitas, Topik Artikel Ilmiah, Publikasi, Sitasi

Abstract

Researcher clustering often uses textual information contained in scientific articles, for example titles, abstracts, and keywords, resulting in groups of researchers with similar textual information from their scientific articles. Researchers clustering also often uses the number of publications and citations, resulting in groups of researchers who tend to have the same number of publications and citations. Based on the two methods above, this study attempts to analyze the use of scientific article topics in the productivity feature extraction process. This feature is a feature obtained through calculating the performance of researchers based on the number of publications and citations. The results of feature extraction will be used for clustering researchers using the K-Means++ method. Before it is clustered, the researcher data must be analyzed first to eliminate the possibility of outliers. Evaluation of cluster results is carried out by considering the sum squared error and silhouette values. As a result, the optimal cluster is obtained with a K value equal to 8 and a silhouette value equal to 0.15396. Then, the results of the clusters are analyzed to be able to label each cluster by considering the topic of scientific articles, number of publications and number of citations.

Keywords: Researcher clustering, Productivity, Topics of Scientific Article, Publication, Citation

1. PENDAHULUAN

Peneliti yang memiliki kinerja publikasi penelitian yang bagus memberikan dampak yang signifikan dalam perkembangan suatu bidang penelitian karena peneliti tersebut memiliki banyak pengalaman dalam mengeksplorasi bidang yang menjadi fokus peneliti. Dalam aktivitas penelitian, peneliti tersebut dapat dijadikan sebagai teman kolaborasi yang sesuai dengan fokus bidang peneliti [1] atau bahkan dengan fokus bidang yang agak berbeda [2]. Ada beberapa metode yang dapat digunakan untuk mengukur kinerja publikasi penelitian seorang peneliti. Pengukuran kinerja publikasi peneliti dapat dilakukan dengan menghitung jumlah artikel ilmiah yang dipublikasikan dan jumlah sitasi yang didapatkan oleh peneliti [3]. Metode pengukuran kinerja publikasi juga dapat menggunakan nilai *impact factor* yang merupakan hasil penghitungan rata-rata total sitasi yang didapatkan dari semua artikel ilmiah yang dipublikasikan oleh peneliti [4]. Semakin banyak sitasi yang didapatkan, semakin tinggi nilai *impact factor* yang dihasilkan. Namun, kekurangan dari nilai *impact factor* adalah tidak memperhitungkan artikel ilmiah yang sedikit atau bahkan tidak mendapatkan sitasi. Untuk mengatasi kekurangan ini, digunakan metode pengukuran kinerja publikasi menggunakan nilai *h-index* [5], yaitu nilai yang merepresentasikan jumlah publikasi penelitian dan jumlah sitasi per publikasi. Sebagai contoh, seorang peneliti memiliki nilai *h-index* sebesar 10 jika telah mempublikasikan minimal 10 artikel ilmiah yang masing-masing artikel ilmiah mendapatkan sitasi sebanyak 10 kali.

Kinerja publikasi peneliti juga dapat digunakan untuk mendapatkan informasi penting yang tersembunyi dari kumpulan peneliti yang ada. Salah satunya adalah informasi tentang klaster peneliti yang memiliki karakteristik tertentu. Untuk itu, klasterisasi peneliti menjadi hal yang perlu dilakukan. Penggunaan fitur yang merupakan hasil transformasi karakteristik peneliti sangat mempengaruhi hasil klasterisasi peneliti. Pada penelitian terdahulu, klasterisasi peneliti dilakukan dengan menggunakan fitur jumlah publikasi [6] sehingga dihasilkan klaster peneliti yang sangat rutin melakukan publikasi sampai klaster peneliti yang sangat jarang melakukan publikasi. Jumlah sitasi yang didapatkan peneliti juga dapat digunakan sebagai fitur untuk menghasilkan klaster peneliti berdasarkan jumlah sitasi yang didapatkan [7]. Pada penelitian lain, data publikasi dan sitasi digunakan untuk menghasilkan fitur produktivitas dan dinamitas [8]. Fitur produktivitas digunakan untuk menghitung jumlah publikasi dan sitasi peneliti setiap tahunnya. Fitur ini juga digunakan untuk mengangkap peningkatan jumlah publikasi dan sitasi peneliti setiap tahunnya. Sementara itu, fitur dinamitas digunakan untuk menangkap kedinamisan perubahan nilai dari fitur produktivitas. Selain fitur yang dihasilkan dari kinerja publikasi peneliti, informasi tekstual yang didapatkan dari judul dan abstrak [9] atau dari kata kunci [10] artikel ilmiah juga bisa digunakan untuk melakukan klasterisasi peneliti.

Selain untuk klasterisasi peneliti, salah satu penggunaan informasi tekstual artikel ilmiah adalah untuk mengekstrak topik penelitian. Kata-kata yang didapatkan dari informasi tekstual diklasterisasi untuk menghasilkan klaster topik penelitian. Salah satu penelitian terdahulu telah melakukan klasterisasi kata-kata yang didapatkan dari informasi tekstual artikel ilmiah menjadi beberapa klaster topik dengan menggunakan metode K-Means [11]. Pada penelitian lainnya, dilakukan klasterisasi kata-kata menggunakan metode *Latent Dirichlet Allocation* (LDA) [12] dengan sumber data artikel ilmiah peneliti. LDA menggunakan informasi kemunculan kata dan panjang informasi tekstual dari artikel ilmiah untuk menghitung distribusi probabilitas topik pada setiap artikel ilmiah. Namun, semakin banyak data yang diolah, semakin lama waktu komputasi yang dibutuhkan.

Pada penelitian ini, diusulkan penggunaan topik artikel ilmiah dalam proses ekstraksi fitur untuk mendapatkan fitur terkait kinerja publikasi peneliti berdasarkan topik artikel ilmiah. Kemudian, hasil ekstraksi fitur digunakan untuk melakukan klasterisasi peneliti sehingga dihasilkan klaster peneliti yang memiliki similaritas dalam hal kinerja publikasi pada topik penelitian tertentu. Kemudian, dilakukan pelabelan secara manual terhadap hasil klaster dengan menganalisis fitur-fitur yang berpengaruh untuk mendeskripsikan hasil klaster.

2. METODE PENELITIAN

Penelitian ini menggunakan data peneliti dari Indonesia untuk dijadikan rekomendasi referensi terhadap penelitian terkait klasterisasi peneliti di Indonesia. Data peneliti yang digunakan berjumlah 4.329 peneliti dengan total publikasi berjumlah 209.184 artikel ilmiah. Pengambilan data peneliti dilakukan secara manual. Sementara itu, data artikel ilmiah peneliti didapatkan dari *Google Scholar* dengan menggunakan *scraping data*. Data artikel ilmiah yang diambil adalah data artikel yang dipublikasikan antara tahun 2010 sampai 2020. Detil data peneliti adalah sebagai berikut.

1. Data-1: 3.894 peneliti yang mendapatkan pendanaan dari Kemenristek dalam 3 tahun terakhir dengan total publikasi berjumlah 181.326 artikel ilmiah
2. Data-2: 500 peneliti teratas jurnal *Science and Technology Index* (SINTA) yang sebagian besar (435 peneliti) tidak termasuk dalam Data-1 dengan total publikasi berjumlah 27.858 artikel ilmiah

2.1 Persiapan data

Sebelum data digunakan, seleksi data secara manual dilakukan untuk menghilangkan data-data peneliti yang memiliki profil belum lengkap atau tidak pernah melakukan publikasi. Proses ini menghasilkan data peneliti berjumlah 3.847 peneliti dengan total publikasi berjumlah 178.911 artikel ilmiah. Kemudian, klasterisasi topik dilakukan terhadap data judul artikel ilmiah berbahasa Indonesia maupun berbahasa Inggris menggunakan metode *deepLDA* [13] yang memiliki waktu komputasi lebih cepat daripada LDA. Proses ini menghasilkan 18 topik yang dapat dilihat pada Tabel 1.

Tabel 1 Hasil ekstraksi topik dan contoh judul artikel ilmiah dari masing-masing topik

Topik	Nama Topik	Contoh Judul Artikel Ilmiah
1	Kimia	Blue Methylene Retrieval using Silica-Salicylic Acid Modified Filtering
2	Hukum	Analisis Pengaruh Kebijakan Hutang Dan Profitabilitas Terhadap Harga Saham Dengan Kebijakan Dividen Dan Nilai Perusahaan Sebagai Variabel Mediasi
3	Nutrisi Boteknologi	Extraction, identification and quantitative HPLC analysis of flavonoids from sarang semut (Myrmecodia pendan)
4	Pertanian	Application of mushroom waste medium from Pleurotus ostreatus for bioremediation of DDT-contaminated soil
5	Mikrobiologi	Efficacy of the new neuraminidase inhibitor CS-8958 against H5N1 influenza viruses
6	Electrical Engineering	Electrochemical energy storage devices for wearable technology: a rationale for materials selection and cell design
7	Pendidikan	Pembelajaran pengantar fisika kuantum dengan memanfaatkan media phet simulation dan LKM melalui pendekatan saintifik: Dampak pada Minat dan Penguasaan Konsep Mahasiswa
8	Lingkungan	Kearifan lokal tentang mitigasi bencana pada masyarakat Baduy
9	Elektro Mesin	A Wide-Speed High Torque Capability Utilizing Overmodulation Strategy in DTC of Induction Machines With Constant Switching Frequency Controller
10	Pendidikan Sains	Exploring agency beyond humans: the compatibility of Actor-Network Theory (ANT) and resilience thinking
11	HRD	The Impact of total quality management practices towards competitive advantage and organizational performance: Case of fishery industry in South Sulawesi Province of Indonesia
12	Peternakan	Effects of meat preheating and wrapping on the levels of polycyclic aromatic hydrocarbons in charcoal-grilled meat
13	Machine Learning	A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing
14	Konservasi	Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (Pongo pygmaeus)
15	Ekonomi Pemberdayaan Masyarakat	Analisis Pemberdayaan Masyarakat Nelayan di Kecamatan Tobelo Kabupaten Halmahera Utara
16	Kimia Organik	Pengaruh Pemberian Probiotik Berbeda pada Pakan Komersial terhadap Pertumbuhan Dan Efisiensi Pakan Ikan Lele Sangkuriang (Clarias Sp.)
17	Penyakit	HIV/AIDS-related mortality in Africa and Asia: evidence from INDEPTH health and demographic surveillance system sites
18	Kesehatan Masyarakat	Efektifitas audiovisual sebagai media penyuluhan kesehatan terhadap peningkatan pengetahuan dan sikap ibu dalam tatalaksana balita dengan diare di dua rumah sakit kota Malang

2.2 Ekstraksi fitur

Pada proses ekstraksi fitur, ditambahkan topik artikel ilmiah yang telah didapatkan sebelumnya sehingga ekstraksi fitur akan menghasilkan 2 jenis fitur terkait kinerja penelitian peneliti pada setiap topik. Penghitungan nilai fitur dilakukan dengan tahun pengujian (t_n) dari 2010 sampai 2020. Fitur produktivitas (F_P) adalah fitur (F_1 sampai F_6) yang dihitung menggunakan kinerja publikasi dan sitasi peneliti p_x terhadap topik (c_a). Sementara itu, fitur dinamitas (F_D) adalah fitur yang menangkap aspek dinamis dari masing-masing fitur produktivitas seperti perubahan minimum, perubahan maksimum, perubahan terakhir, total perubahan, dan nilai representatif kinerja keseluruhan peneliti. Detil fitur dapat dilihat pada Tabel 2.

Tabel 2 Fitur Produktivitas dan Dinamitas Peneliti

Fitur	Jenis Fitur	Deskripsi
$F_1(p_x, c_a, t_n)$	Produktivitas	Jumlah artikel penelitian dengan topik c_a yang dipublikasikan peneliti p_x pada tahun t_n
$F_2(p_x, c_a, t_m, t_n)$	Produktivitas	Jumlah kumulatif artikel penelitian dengan topik c_a yang dipublikasikan peneliti p_x dari tahun t_m sampai tahun t_n
$F_3(p_x, c_a, t_m, t_n)$	Produktivitas	Jumlah kumulatif bobot artikel penelitian dengan topik c_a yang dipublikasikan peneliti p_x dari tahun t_m sampai tahun t_n yang dihitung menggunakan persamaan $\sum_{t_o=t_m}^{t_n} \frac{F_1(p_x, c_a, t_o)}{t_n - t_o + 1}$
$F_4(p_x, c_a, t_n)$	Produktivitas	Jumlah sitasi yang mengarah pada artikel penelitian dengan topik c_a milik peneliti p_x yang didapatkan pada tahun t_n
$F_5(p_x, c_a, t_m, t_n)$	Produktivitas	Jumlah kumulatif sitasi yang mengarah pada artikel penelitian dengan topik c_a milik peneliti p_x dari tahun t_m sampai tahun t_n
$F_6(p_x, c_a, t_m, t_n)$	Produktivitas	Jumlah kumulatif bobot sitasi yang mengarah pada artikel penelitian dengan topik c_a milik peneliti p_x dari tahun t_m sampai tahun t_n yang dihitung menggunakan persamaan $\sum_{t_o=t_m}^{t_n} \frac{F_4(p_x, c_a, t_o)}{t_n - t_o + 1}$
$F_7(F_P, p_x, c_a, t_m, t_n)$	Dinamitas	Selisih minimal terhadap perubahan nilai fitur F_P dengan topik c_a yang dimiliki peneliti p_x dari tahun t_m sampai tahun t_n yang dihitung menggunakan persamaan $\min_{t_{m+1} \ldots t_n} (F_P(p_x, c_a, t_o) - F_P(p_x, c_a, t_{o-1}))$
$F_8(F_P, p_x, c_a, t_m, t_n)$	Dinamitas	Selisih maksimal terhadap perubahan nilai fitur F_P dengan topik c_a yang dimiliki peneliti p_x dari tahun t_m sampai tahun t_n yang dihitung menggunakan persamaan $\max_{t_{m+1} \ldots t_n} (F_P(p_x, c_a, t_o) - F_P(p_x, c_a, t_{o-1}))$
$F_9(F_P, p_x, c_a, t_m, t_n)$	Dinamitas	Total perubahan nilai fitur F_P dengan topik c_a yang dimiliki peneliti p_x dari tahun t_m sampai tahun t_n yang dihitung menggunakan persamaan $\sum_{t_o=t_{m+1}}^{t_n} (F_P(p_x, c_a, t_o) - F_P(p_x, c_a, t_{o-1}))$
$F_{10}(F_P, p_x, c_a, t_n)$	Dinamitas	Perubahan terakhir nilai fitur F_P dengan topik c_a yang dimiliki peneliti p_x pada tahun t_n yang dihitung menggunakan persamaan $F_P(p_x, c_a, t_n) - F_P(p_x, c_a, t_{n-1})$
$F_{11}(F_P, p_x, c_a, t_m, t_n)$	Dinamitas	Nilai representatif fitur F_P dengan topik c_a yang dimiliki peneliti p_x dari tahun t_m sampai tahun t_n

Untuk menghitung F_{11} dari F_1 dan F_4, digunakan persamaan (1). Untuk menghitung F_{11} dari F_2 dan F_5, digunakan persamaan (2). Untuk menghitung F_{11} dari F_3 dan F_6, digunakan persamaan (3). Perpaduan 6 fitur produktivitas (F_P), 18 topik artikel ilmiah (c_a), dan 11 tahun pengujian (t_n) menghasilkan 1.188 fitur (F_P x c_a x t_n). Kemudian, perpaduan 5 fitur dinamitas (F_D), 6 fitur produktivitas (F_P), dan 18 topik artikel ilmiah (c_a) menghasilkan 540 fitur (F_P x F_D x c_a). Jadi, total keseluruhan fitur yang digunakan untuk klasterisasi peneliti berjumlah 1.728 fitur.

$$\sum_{t_o=t_m}^{t_n} \frac{F_P(p_x, c_a, t_o)}{n} \quad (1)$$

$$\frac{F_P(p_x, c_a, t_n)}{n} \quad (2)$$

$$F_P(p_x, c_a, t_n) \quad (3)$$

Tabel 3 Contoh data publikasi dan sitasi peneliti p_x

Artikel Ilmiah	Tahun Publikasi	Tahun Sitasi			
		2018	2019	2020	2021
ar_1	2018	2	2	0	0
ar_2	2018	0	3	0	0
ar_3	2019		0	4	0
ar_4	2020			0	0
ar_5	2020			0	0
ar_6	2020			0	5
ar_7	2021				0
ar_8	2021				0

Pada Tabel 3, ditunjukkan contoh data publikasi dan sitasi peneliti p_x yang melakukan publikasi artikel ilmiah ar_1 sampai ar_8 dengan topik c_a antara tahun 2018 sampai 2021. Dari data di atas, dilakukan penghitungan fitur peneliti dengan tahun pengujian t_n sama dengan 2020. Untuk penghitungan fitur F_7 sampai F_{11}, digunakan fitur F_1 sebagai acuan. Hasil penghitungan fitur dapat dilihat pada Tabel 4.

Tabel 4 Contoh penghitungan fitur peneliti

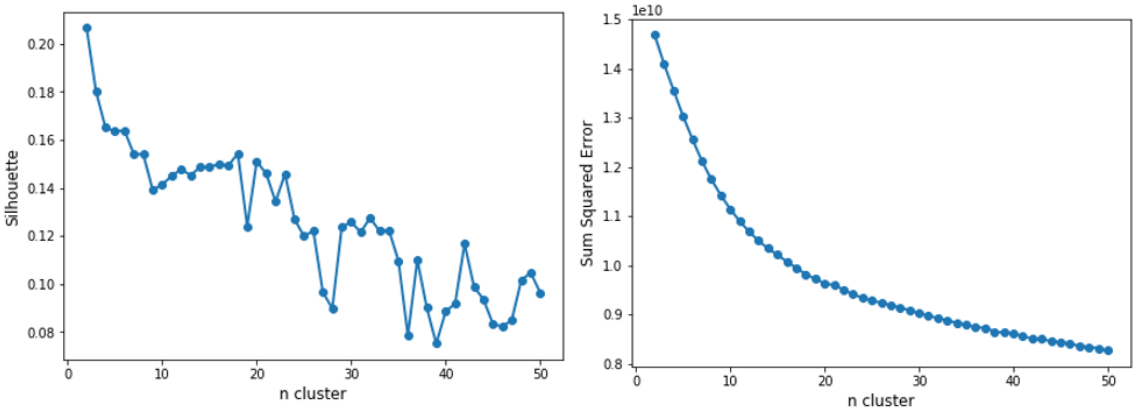
Fitur	Nilai Penghitungan Fitur
$F_1(p_x, c_a, t_n)$	$F_1(p_x, c_a, 2020) = 3$
$F_2(p_x, c_a, t_m, t_n)$	$F_2(p_x, c_a, 2018{,}2020) = 2 + 1 + 3 = 6$
$F_3(p_x, c_a, t_m, t_n)$	$\sum_{t_o=2018}^{2020} \frac{F_1(p_x,c_a,t_o)}{2020-t_o+1} = \left(\frac{F_1(p_x,c_a,2018)}{2020-2018+1}\right) + \left(\frac{F_1(p_x,c_a,2019)}{2020-2019+1}\right) + \left(\frac{F_1(p_x,c_a,2020)}{2020-2020+1}\right) = \frac{2}{3} + \frac{1}{2} + \frac{3}{1} = 4.17$
$F_4(p_x, c_a, t_n)$	$F_4(p_x, c_a, 2020) = 4$
$F_5(p_x, c_a, t_m, t_n)$	$F_5(p_x, c_a, 2018{,}2020) = (2 + 0) + (2 + 3 + 0) + (0 + 0 + 4 + 0 + 0 + 0) = 2 + 5 + 4 = 11$
$F_6(p_x, c_a, t_m, t_n)$	$\sum_{t_o=2018}^{2020} \frac{F_4(p_x,c_a,t_o)}{2020-t_o+1} = \left(\frac{F_4(p_x,c_a,2018)}{2020-2018+1}\right) + \left(\frac{F_4(p_x,c_a,2019)}{2020-2019+1}\right) + \left(\frac{F_4(p_x,c_a,2020)}{2020-2020+1}\right) = \frac{2}{3} + \frac{5}{2} + \frac{4}{1} = 7.17$
$F_7(F_P, p_x, c_a, t_m, t_n)$	$\min_{2019\ldots2020}\left(F_1(p_x, c_a, t_o) - F_1(p_x, c_a, t_{o-1})\right) = \min\Big(\big(F_1(p_x, c_a, 2019) - F_1(p_x, c_a, 2018)\big), \big(F_1(p_x, c_a, 2020) - F_1(p_x, c_a, 2019)\big)\Big) = \min\big((1-2), (3-1)\big) = \min(-1, 2) = -1$
$F_8(F_P, p_x, c_a, t_m, t_n)$	$\max_{2019\ldots2020}\left(F_1(p_x, c_a, t_o) - F_1(p_x, c_a, t_{o-1})\right) = \max\Big(\big(F_1(p_x, c_a, 2019) - F_1(p_x, c_a, 2018)\big), \big(F_1(p_x, c_a, 2020) - F_1(p_x, c_a, 2019)\big)\Big) = \max\big((1-2), (3-1)\big) = \max(-1, 2) = 2$
$F_9(F_P, p_x, c_a, t_m, t_n)$	$\sum_{t_o=2019}^{t_n}\left(F_1(p_x, c_a, t_o) - F_1(p_x, c_a, t_{o-1})\right) = \big(F_1(p_x, c_a, 2019) - F_1(p_x, c_a, 2018)\big) + \big(F_1(p_x, c_a, 2020) - F_1(p_x, c_a, 2019)\big) = (1-2) + (3-1) = -1 + 2 = 1$
$F_{10}(F_P, p_x, c_a, t_n)$	$F_1(p_x, c_a, 2020) - F_1(p_x, c_a, 2019) = 3 - 1 = 2$
$F_{11}(F_P, p_x, c_a, t_m, t_n)$	$\sum_{t_o=2018}^{2020} \frac{F_1(p_x,c_a,t_o)}{2020} = \left(\frac{F_1(p_x,c_a,2018)}{2020}\right) + \left(\frac{F_1(p_x,c_a,2019)}{2020}\right) + \left(\frac{F_1(p_x,c_a,2020)}{2020}\right) = \frac{2}{2020} + \frac{1}{2020} + \frac{3}{2020} = 0.003$

2.3 Klasterisasi menggunakan metode K-Means

Algoritma K-Means adalah algoritma pengklasteran yang diusulkan oleh J.B. MacQueen pada tahun 1967. K-Means merupakan algoritma unsupervised [14] sehingga dalam proses klasterisasinya tidak menggunakan model pembelajaran yang dihasilkan dari pelatihan terhadap data *training*. Algoritma K-Means akan mengklaster data menjadi k klaster berdasarkan jarak terdekat setiap data dengan k *centroid* yang umumnya dipilih secara acak. Algoritma ini sangat dipengaruhi oleh metode perhitungan yang digunakan untuk menghitung jarak data dengan *centroid*. Metode yang umum digunakan untuk menghitung jarak adalah *Euclidean distance*. Pada penelitian ini, klasterisasi peneliti dilakukan dengan menggunakan dengan nilai K = 2 sampai K = 50. Data input yang digunakan berdimensi 3834 baris (jumlah peneliti) dan 1728 kolom (jumlah fitur).

3. HASIL DAN PEMBAHASAN

Untuk menganalisis hasil klasterisasi, dilakukan 2 skenario uji coba. Yang pertama, menentukan nilai K untuk menghasilkan klaster peneliti yang optimal. Yang kedua, menganalisis hasil klaster untuk menentukan label terhadap hasil klasterisasi peneliti. Penghapusan *outlier* [15] perlu dilakukan karena hasil klasterisasi awal menunjukkan distribusi data peneliti yang sangat tidak merata dengan nilai K = 8 menghasilkan nilai silhouette sebesar 0.92816. Sebanyak 3834 peneliti terdapat pada klaster 1, sementara 13 peneliti sisanya terbagi ke dalam 7 klaster lain. Untuk mendapatkan hasil klasterisasi dengan distribusi data peneliti yang lebih merata, data 13 peneliti tersebut dihapus sehingga penelitian ini menggunakan data 3834 peneliti dengan 178.118 artikel ilmiah.



Gambar 1 Nilai SSE dan Silhouette Hasil Klasterisasi Peneliliti

3.2 Menentukan nilai K Optimal

Untuk menentukan nilai K yang optimal terhadap klasterisasi peneliti, digunakan nilai *sum squared error* (SSE) dan *silhouette*. SSE digunakan untuk menguji validitas hasil klasterisasi dengan menentukan *error* dari setiap data dengan titik tengahnya [16]. Sementara itu, *silhouette* membandingkan nilai similaritas data di dalam suatu klaster dengan nilai similaritas data klaster lain untuk mengevaluasi konsistensi data pada hasil klasterisasi [19,20]. Berdasarkan Gambar 1, semakin bertambah nilai K, nilai SSE cenderung menurun. Ini karena semakin tinggi nilai K, semakin kecil ruang lingkup klaster yang dihasilkan. Dengan menggunakan metode *elbow* [19], berdasarkan nilai SSE, kemungkinan nilai K optimal berada pada nilai K = 8 sampai K = 17. Hal ini karena nilai K = 2 sampai K = 7 menghasilkan penurunan nilai SSE yang cukup signifikan dan nilai K = 17 ke atas menghasilkan penurunan nilai SSE yang kecil.

Setelah ditemukan rentang nilai K yang diprediksi menghasilkan klaster optimal dari analisis sebelumnya, digunakan nilai *silhouette* untuk menentukan satu nilai K yang paling optimal. Dari nilai K = 8 sampai K = 17, nilai *silhouette* tertinggi didapatkan dengan nilai K = 8 sebesar 0.15396. Jadi, nilai K optimal untuk klasterisasi peneliti adalah K = 8.

3.3 Analisis untuk pelabelan hasil klasterisasi peneliti

Untuk melakukan pelabelan terhadap hasil klasterisasi peneliti, diperlukan analisis untuk menentukan kecenderungan klaster terhadap topik artikel ilmiah. Asumsi awal kecenderungan klaster dianalisis dengan menghitung rata-rata jumlah publikasi per peneliti (M_P) dan rata-rata jumlah sitasi per peneliti (M_S) setiap klaster pada masing-masing topik artikel ilmiah. Jumlah publikasi dan jumlah sitasi digunakan karena 2 aspek inilah yang menjadi acuan dasar dalam melakukan penghitungan F_1 sampai F_{11}. Dari hasil penghitungan M_P dan M_S pada Tabel 5 dan Tabel 6, didapatkan topik dengan nilai M_P tertinggi dan topik dengan nilai M_S tertinggi sebagai topik utama dari masing-masing klaster.

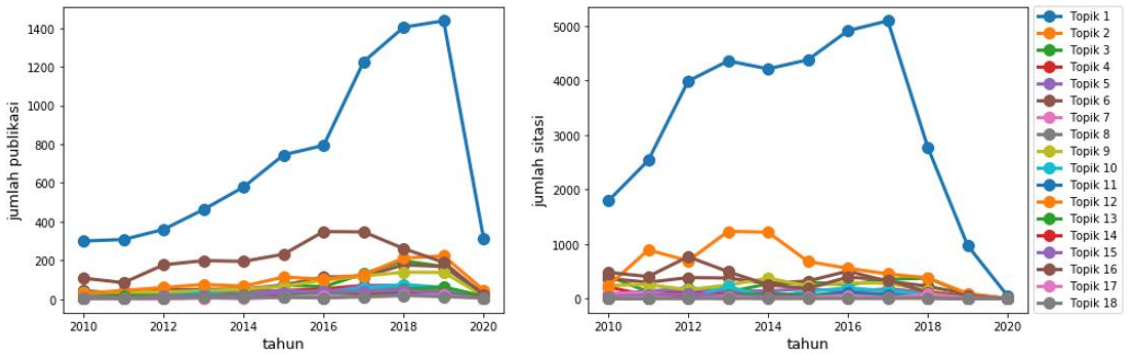
Tabel 5 Hasil penghitungan rata-rata jumlah publikasi per peneliti di setiap klaster

K	Topik																	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	26	1	3	1	1	3	1	1	3	1	1	4	1	0	1	7	0	0
2	1	3	1	2	1	1	2	9	1	2	3	2	2	19	14	1	1	1
3	0	1	2	0	6	1	2	0	1	3	2	1	1	2	1	2	21	13
4	3	1	5	15	3	2	1	2	1	1	1	14	1	3	2	6	1	0
5	2	3	0	1	1	8	2	5	14	4	2	2	14	1	1	1	1	1
6	1	2	1	1	1	1	4	2	1	3	2	1	1	2	2	2	2	2
7	3	1	15	3	12	2	1	0	1	1	0	3	1	2	2	12	3	2
8	0	15	0	0	0	1	4	2	2	4	19	0	3	2	3	0	1	2

Tabel 6 Hasil penghitungan nilai rata-rata jumlah sitasi per peneliti di setiap klaster

K	Topik																	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	115	1	8	3	4	10	2	1	8	3	2	21	1	1	1	11	1	0
2	1	3	4	4	6	3	4	14	5	6	14	5	7	88	28	3	5	4
3	1	1	9	1	24	4	5	0	2	7	6	1	2	5	2	4	68	31
4	13	1	16	37	13	5	2	3	3	1	2	55	2	12	4	16	2	2
5	6	4	1	1	1	22	7	4	47	11	6	6	45	4	1	1	2	1
6	2	2	2	1	3	2	11	1	2	6	3	2	3	3	2	2	2	3
7	8	1	53	8	49	7	3	1	1	2	1	9	3	4	3	26	5	3
8	0	62	1	5	0	4	13	3	4	14	58	2	8	4	7	1	2	5

Berdasarakan Tabel 7, penghitungan nilai M_P dan M_S menghasilkan topik utama 1 (TU-1) yang sama hampir pada setiap klaster. Klaster 1 menghasilkan TU-1 yang sama (Kimia) dari penghitungan nilai M_P dan M_S. Hal ini juga dapat dilihat pada klaster 2, 3, 5, 6, 7 yang masing-masing menghasilkan TU-1 yang sama dari penghitungan nilai M_P dan M_S. Namun, penghitungan nilai M_P dan M_S menghasilkan TU-1 yang berbeda pada klaster 4 dan 8. Klaster 4 menghasilkan TU-1 Pertanian dan TU-2 Peternakan dari penghitungan nilai M_P. Kemudian dari penghitungan nilai M_S, klaster 4 menghasilkan TU-1 Peternakan dan TU-2 Pertanian. Jika dianalisis lebih lanjut, topik Pertanian dan Peternakan ternyata merupakan 2 topik teratas dari masing-masing penghitungan nilai M_P dan M_S pada klaster 4. Kasus ini juga terjadi pada klaster 8. Setelah diketahui kecenderungan topik pada setiap klaster dengan menghitung nilai M_P dan M_S, validasi hasil klaster dilakukan dengan menganalisis fitur-fitur peneliti pada klaster 1.



Gambar 2 Jumlah publikasi dan sitasi peneliti klaster 1

Berdasarkan Gambar 2, jumlah publikasi (F_1) keseluruhan dari semua peneliti setiap tahun pada topik 1 (Kimia) jauh lebih banyak dari jumlah publikasi pada topik lainnya. Begitu juga jumlah sitasi (F_4) keseluruhan setiap tahun pada topik 1 (Kimia) yang jauh lebih banyak dari jumlah sitasi pada topik lainnya. Jika dianalisis lebih lanjut, sebanyak 78% peneliti dari klaster 1 memiliki jumlah publikasi paling banyak pada topik 1 (Kimia). Dari sisi jumlah sitasi, persentase peneliti klaster 1 yang mendapatkan sitasi paling banyak pada topik 1 (Kimia) juga sebanyak 78% peneliti. Selanjutnya, dari fitur F_2 dan F_5 yang merupakan nilai akumulatif

jumlah publikasi dan sitasi setiap tahunnya, didapatkan laju peningkatan jumlah publikasi dan sitasi peneliti. Dari hasil analisis, didapatkan sebanyak 79% peneliti memiliki laju peningkatan publikasi terbaik dan 78% peneliti memiliki laju peningkatan sitasi terbaik pada topik 1 (Kimia).

Tabel 7 Topik utama pada masing-masing klaster

Klaster	Jumlah Peneliti	Rata-rata jumlah publikasi per peneliti (M_P)	Rata-rata jumlah sitasi per peneliti (M_P)
1	306	TU-1: Topik 1 (Kimia) = 26 TU-2: Topik 16 (Kimia organik) = 7 TU-3: Topik 12 (Peternakan) = 4	TU-1: Topik 1 (Kimia) = 115 TU-2: Topik 12 (Peternakan) = 21 TU-3: Topik 16 (Kimia Organik) = 11
2	276	TU-1: Topik 14 (Konservasi) = 19 TU-2: Topik 15 (Ekonomi Pemberdayaan Masyarakat) = 14 TU-3: Topik 8 (Lingkungan) = 9	TU-1: Topik 14 (Konservasi) = 88 TU-2: Topik 15 (Ekonomi Pemberdayaan Masyarakat) = 28 TU-3: Topik 8 (Lingkungan) = 14, Topik 11 (HRD) = 14
3	263	TU-1: Topik 17 (Penyakit) = 21 TU-2: Topik 18 (Kesehatan Masyarakat) = 13 TU-3: Topik 5 (Mikrobiologi) = 6	TU-1: Topik 17 (Penyakit) = 68 TU-2: Topik 18 (Kesehatan Masyarakat) = 31 TU-3: Topik 5 (Mikrobiologi) = 24
4	268	TU-1: Topik 4 (Pertanian) = 15 TU-2: Topik 12 (Peternakan) = 14 TU-3: Topik 16 (Kimia Organik) = 6	TU-1: Topik 12 (Peternakan) = 55 TU-2: Topik 4 (Pertanian) = 37 TU-3: Topik 3 (Nutrisi Bioteknologi) = 16, Topik 16 (Kimia Organik) = 16
5	344	TU-1: Topik 9 (Elektro Mesin) = 14 Topik 13 (Machine Learning) = 14 TU-2: Topik 6 (Electrical Engineering) = 8 TU-3: Topik 8 (Lingkungan) = 5	TU-1: Topik 9 (Elektro Mesin) = 47 TU-2: Topik 13 (Machine Learning) = 45 TU-3: Topik 6 (Electrical Engineering) = 22
6	1736	TU-1: Topik 7 (Pendidikan) = 4 TU-2: Topik 10 (Pendidikan Sains) = 3 TU-3: Topik 2 (Hukum) = 2, Topik 8 (Lingkungan) = 2, Topik 11 (HRD) = 2, Topik 14 (Konservasi) = 2, Topik 15 (Ekonomi Permberdayaan Masyarakat) = 2, Topik 16 (Kimia Organik) = 2, Topik 17 (Penyakit) = 2, Topik 18 (Kesehatan Masyarakat) = 2	TU-1: Topik 7 (Pendidikan) = 11 TU-2: Topik 10 (Pendidikan Sains) = 6 TU-3: Topik 12 (Peternakan) = 4 Topik 5 (Mikrobiologi) = 3, Topik 11 (HRD) = 3, Topik 13 (Machine Learning) = 3, Topik 14 (Konservasi) = 3, Topik 18 (Kesehatan Masyarakat) = 3
7	363	TU-1: Topik 3 (Nutrisi Bioteknologi) = 15 TU-2: Topik 5 (Mikrobiologi) = 12, Topik 16 (Kimia Organik) = 12 TU-3: Topik 1 (Kimia) = 3, Topik 4 (Pertanian) = 3, Topik 17 (Penyakit) = 3	TU-1: Topik 3 (Nutrisi Bioteknologi) = 53 TU-2: Topik 5 (Mikrobiologi) = 49 TU-3: Topik 16 (Kimia Organik) = 26
8	278	TU-1: Topik 11 (HRD) = 19 TU-2: Topik 2 (Hukum) = 15 TU-3: Topik 7 (Pendidikan) = 4, Topik 10 (Pendidikan Sains) = 4	TU-1: Topik 2 (Hukum) = 62 TU-2: Topik 11 (HRD) = 58 TU-3: Topik 10 (Pendidikan Sains) = 14

Tabel 8 Persentase jumlah peneliti dengan nilai perubahan terbaik pada topik 1 (Kimia)

Fitur Produktivitas	Fitur Dinamis				
	F_7	F_8	F_9	F_{10}	F_{11}
F_1	72%	74%	60%	15%	78%
F_4	72%	70%	83%	58%	78%

Aspek dinamis digunakan untuk menghitung persentase jumlah peneliti dengan perubahan tertinggi fitur jumlah publikasi (F_1) dan jumlah sitasi (F_4) dengan melihat perubahan minimum (F_7), perubahan maksimum (F_8), perubahan total (F_9), perubahan terakhir (F_{10}), dan nilai representatif (F_{11}). Fitur F_2, F_3, F_5, dan F_6 merupakan fitur yang menghasilkan nilai akumulasi dari tahun sebelumnya sehingga tidak bisa digunakan dalam analisis aspek dinamis. Berdasarkan Tabel 8, nilai perubahan F_1 dan F_4 pada topik 1 (Kimia) menghasilkan persentase jumlah peneliti yang tinggi pada setiap fitur dinamis (F_D). Namun, kasus berbeda ditunjukkan

F_{10} terhadap F_1 yang menghasilkan 15% peneliti dengan nilai F_{10} terbaik. Jika dianalisis pada Gambar 2, publikasi artikel ilmiah dengan topik 1 (Kimia) mengalami penurunan yang signifikan dari tahun 2019 sebanyak 1437 artikel ilmiah menjadi 312 artikel ilmiah di tahun 2020. Ini menunjukkan sebagian besar peneliti klaster 1 mengalami penurunan yang signifikan terkait jumlah publikasi dari tahun 2019 ke tahun 2020 yang merupakan perubahan terakhir F_1. Dari analisis beberapa fitur yang telah dilakukan, dapat disimpulkan bahwa sebagian besar peneliti klaster 1 memiliki kecenderungan yang sangat tinggi terhadap topik 1 (Kimia). Dengan menggunakan informasi yang dihasilkan dari Tabel 7, dilakukan pelabelan terhadap hasil klaster sebagai berikut.

- Klaster 1 merupakan klaster peneliti yang memiliki kecenderungan sangat tinggi pada bidang kimia dengan jumlah publikasi dan jumlah sitasi sangat dominan daripada bidang yang lain.
- Klaster 2 merupakan klaster peneliti yang memberikan dampak signifikan terhadap bidang konservasi. Walaupun jumlah publikasi mulai mengalami kenaikan yang signifikan pada 2017, tetapi jumlah sitasi tertinggi justru didapatkan pada tahun 2015.
- Klaster 3 merupakan klaster peneliti yang memiliki kecenderungan sangat tinggi pada bidang kesehatan dengan jumlah publikasi dan jumlah sitasi sangat dominan daripada bidang yang lain.
- Klaster 4 merupakan klaster peneliti yang memiliki kecenderungan 2 bidang yang hampir berimbang, yaitu bidang pertanian dan peternakan. Dari sisi publikasi, bidang pertanian menghasilkan publikasi yang lebih banyak daripada bidang peternakan. Namun, dari sisi sitasi bidang peternakan memiliki jumlah sitasi yang lebih unggul setiap tahunnya dari bidang pertanian.
- Klaster 5 merupakan klaster peneliti yang memiliki kecenderungan yang sama terhadap bidang elektro mesin dan *machine learning*. Hal ini bisa dilihat dari nilai jumlah publikasi per peneliti yang dihasilkan dari kedua bidang tersebut.
- Klaster 6 merupakan klaster peneliti dengan jumlah 5 kali lipat dari jumlah peneliti dari klaster lain. Peneliti dari klaster ini menghasilkan publikasi dan sitasi di bidang pendidikan yang sedikit lebih unggul dari bidang-bidang lainnya.
- Klaster 7 merupakan klaster peneliti yang memiliki kecenderungan pada bidang nutrisi bioteknologi. Klaster ini juga memiliki kecenderungan lain pada bidang mikrobiologi dan kimia organik.
- Klaster 8 merupakan klaster peneliti yang memiliki kecenderungan pada bidang HRD dan hukum. Dari sisi jumlah publikasi bidang HRD lebih unggul, tetapi dari sisi jumlah sitasi bidang hukum yang lebih unggul.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan yang dipaparkan di atas, dapat diambil kesimpulan bahwa klasterisasi peneliti dengan jumlah bersih 3.834 peneliti dan 178.118 artikel ilmiah menghasilkan nilai K = 8 sebagai nilai klaster yang optimal dengan nilai *silhouette* sebesar 0.15396. Kemudian, pelabelan hasil klaster dilakukan dengan menganalisis beberapa fitur yang memiliki peran dalam proses klasterisasi. Dari hasil analisis yang dilakukan, didapatkan setiap klaster peneliti yang memiliki kecenderungan topik atau bidang yang berbeda-beda.

Untuk pengembangan ke depannya, penambahan fitur kolaborasi dalam proses ekstraksi fitur bisa dilakukan. Pada proses klasterisasi, bisa ditambahkan beberapa metode sebagai perbandingan untuk menentukan metode terbaik dalam pengklasteran peneliti.

DAFTAR PUSTAKA

[1] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 364–375, 2014, doi: 10.1109/TETC.2014.2356505.

[2] D. Purwitasari, C. Fatichah, I. K. E. Purnama, S. Sumpeno, and M. H. Purnomo, "Inter-departmental research collaboration recommender system based on content filtering in a cold start problem," *2017 IEEE 10th Int. Work. Comput. Intell. Appl. IWCIA 2017 - Proc.*, vol. 2017-Decem, pp. 177–184, 2017, doi: 10.1109/IWCIA.2017.8203581.

[3] M. Ochsner, S. E. Hug, and H. D. Daniel, "Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies," *Zeitschrift fur Erziehungswiss.*, vol. 17, no. 6, pp. 111–132, 2014, doi: 10.1007/s11618-014-0576-4.

[4] P. O. Seglen, "Why the impact factor of journals should not be used for evaluating research," *Br. Med. J.*, vol. 314, no. 7079, pp. 497–497, 1997, doi: 10.1136/bmj.314.7079.497.

[5] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 46, pp. 16569–16572, 2005, doi: 10.1073/pnas.0507655102.

[6] T. Y. Alkan, M. Gunay, and F. Ozbek, "Clustering of Scientist using Research Areas at Akdeniz University," *UBMK 2019 - Proceedings, 4th Int. Conf. Comput. Sci. Eng.*, pp. 580–584, 2019, doi: 10.1109/UBMK.2019.8907236.

[7] D. Yu, W. Wang, S. Zhang, W. Zhang, and R. Liu, "Hybrid self-optimized clustering model based on citation links and textual features to detect research topics," *PLoS One*, vol. 12, no. 10, pp. 1–21, 2017, doi: 10.1371/journal.pone.0187164.

[8] G. Panagopoulos, G. Tsatsaronis, and I. Varlamis, "Detecting rising stars in dynamic collaborative networks," *J. Informetr.*, vol. 11, no. 1, pp. 198–222, 2017, doi: 10.1016/j.joi.2016.11.003.

[9] H. Ma'rifah, A. P. Wibawa, and M. I. Akbar, "Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing," *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 2, no. 2, p. 70, 2020, doi: 10.30872/jsakti.v2i2.2681.

[10] A. M. Rukmi and I. M. Iqbal, "Using k-means++ algorithm for researchers clustering," *AIP Conf. Proc.*, vol. 1867, no. August, 2017, doi: 10.1063/1.4994455.

[11] D. Purwitasari, C. Fatichah, S. Sumpeno, and M. H. Purnomo, "Ekstraksi Ciri Produktivitas Dinamis untuk Prediksi Topik Pakar dengan Model Discrete Choice," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 4, pp. 418–426, 2018, doi: 10.22146/jnteti.v7i4.460.

[12] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, "Metode Latent Dirichlet Allocation untuk Ekstraksi Topik Dokumen," *J. Log.*, vol. 17, no. 3, pp. 154–158, 2017, doi: 10.31940/logic.v17i3.604.

[13] D. Zhang, T. Luo, and D. Wang, "Learning from LDA using deep neural networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10102, pp. 657–664, 2016, doi: 10.1007/978-3-319-50496-4_59.

[14] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012, doi: 10.1016/j.phpro.2012.03.206.

[15] A. Barai (Deb) and L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering," *World J. Comput. Appl. Technol.*, vol. 5, no. 2, pp. 24–29, 2017, doi: 10.13189/wjcat.2017.050202.

[16] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, 2019, doi: 10.1088/1742-6596/1361/1/012015.

[17] F. Wang, H. H. Franco-Penya, J. D. Kelleher, J. Pugh, and R. Ross, "An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10358 LNAI, no. July, pp. 291–305, 2017, doi: 10.1007/978-3-319-62416-7_21.

[18] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, and M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *Int. J. Eng. Technol.*, vol. 7, pp. 105–109, 2018, doi: 10.14419/ijet.v7i2.14.11464.

[19] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.