PERBANDINGAN APLIKASI ALGORITMA KERNEL K-MEANS PADA GRAF BIPARTIT DAN K-MEANS PADA MATRIKS DOKUMEN-ISTILAH DALAM DATASET PENELITIAN COVID-19 RISTEKBRIN

Budi Nugroho*¹

¹Kelompok Penelitian Infometrika dan Informatika Sosial, Puslit Informatika - LIPI

Email: ¹budi.nugroho@lipi.go.id

*Penulis Korespondensi

(Naskah masuk: 17 November 2020, diterima untuk diterbitkan: 22 Maret 2021)

Abstrak

Merebaknya kasus Covid-19 di Indonesia telah memunculkan berbagai macam topik penelitian yang dilakukan oleh para peneliti di berbagai bidang dan dari bermacam institusi. Berdasarkan data yang dihimpun oleh portal Sinta Ristekbrin, terdapat 351 topik penelitian yang telah diunggah oleh para peneliti. Kajian ini dimaksudkan untuk menganalisis dan memetakan topik penelitian yang sedang dan/atau telah dilakukan selama kurun waktu terjadinya pandemi Covid-19 di tanah air. Analisis dan pemetaan dilakukan dengan menerapkan algoritma kernel *k-means* untuk klastering dokumen berbasis graf bipartit dan *k-means* pada matriks dokumen-istilah. Dataset penelitian Covid-19 Ristekbrin dimodelkan sebagai graf bipartit antara daftar istilah dengan dokumennya. Selanjutnya skor kemiripan dihitung dengan metode kernel. Nilai matriks kernel yang mencerminkan skor kemiripan antar dokumen digunakan sebagai masukan bagi algoritma klastering kernel *k-means* yang memberikan hasil berupa pemetaan topik penelitian. Sebagai pembandingan, algoritma *k-means* diterapkan pada matriks dokumen-istilah untuk klastering topik penelitian Covid-19. Dari kedua metode tersebut, hasil klastering diuji dengan validasi internal menggunakan indeks Dunn. Indeks Dunn digunakan karena dalam dataset tidak tersedia informasi awal mengenai label atau nama dari masing-masing klaster. Hasil penelitian ini menunjukkan bahwa algoritma kernel *k-means* memberikan validasi yang sedikit lebih baik dibandingkan dengan *k-means*. Hasil kajian ini diharapkan dapat memberikan tambahan informasi yang mendukung program pemerintah dalam mempercepat penanganan Covid-19 di Indonesia.

Kata kunci: Covid-19, pemetaan, topik penelitian, algoritma kernel *k-means*, graf bipartit

THE APPLICATION OF KERNEL K-MEANS ALGORITHM ON BIPARTITE GRAPH OF THE RISTEKBRIN DATASET OF COVID-19 RESEARCH

Abstract

The outbreak concerning the Covid-19 case in Indonesia has raised various topics of research carried out by researchers in diverse fields and from many institutions. Based on data compiled by the Sinta Ristekbrin portal, 351 research topics have been uploaded by researchers. This study is aimed to analyze and map research topics that are being and/or have been conducted during the period of the Covid-19 pandemic in Indonesia. Analysis and mapping are accomplished by applying the kernel *k-means* algorithm for document clustering based on bipartite graphs and *k-means* on document term matrix. Ristekbrin's Covid-19 research dataset is modeled as a bipartite graph between terms and documents. Furthermore, the similarity score is calculated using the kernel method. The kernel matrix value that represents the similarity score between documents is used as input for the kernel *k-means* clustering algorithm, which provides the results of mapping research topics. As comparison, we applied original *k-means* algorithm on the document-term matrix of the dataset. From these two methods, the clustering results were validated using Dunn index as an internal validation. The Dunn index was used because the dataset did not provide initial information regarding the label or name of each clusters. The comparison Dunn index shows that the kernel *k-means* algorithm outperforms than the *k-means* algorithm. This study is expected to provide additional information that supports government programs in accelerating the handling of Covid-19 in Indonesia.

Keywords: Covid-19, research topic mapping, kernel *k-means* algorithm, bipartite graph

1. PENDAHULUAN

Terjadinya pandemi COVID-19 telah memicu kesadaran para akademisi untuk melakukan penelitian dan pengembangan (litbang) dalam rangka membantu pemerintah dan masyarakat mempercepat penanggulangan dampak dari pandemi ini. Tak terkecuali di Indonesia, pemerintah melalui Kementerian Riset dan Teknologi / Badan Riset dan Inovasi Nasional (Kemenristek/BRIN) mendorong para akademisi di lingkup nasional untuk melakukan litbang terkait dengan Covid-19. Salah satu upaya yang dilakukan adalah dengan menghimpun informasi kegiatan litbang, publikasi ilmiah dan dataset Covid-19 di portal Sinta Kemenristek/BRIN. Portal tersebut dapat diakses melalui alamat URL <http://sinta.ristekbrin.go.id/covid/>. Gambar 1 berikut ini menampilkan tangkapan layar dari situs tersebut.

Gambar. 1 Tangkapan Layar Situs Pangkalan Data Covid-19 Kemenristek/BRIN

Semenjak diumumkannya kasus pertama Covid-19 pada bulan Maret 2020 yang lalu, telah banyak kegiatan litbang yang dilakukan oleh para akademisi dari lembaga litbang dan perguruan tinggi dengan berbagai macam topik yang spesifik. Penelitian Sukmana, Aminuddin dan Nopriyanto mengungkapkan bahwa pandemi Covid-19 telah berdampak di berbagai sektor yaitu kesehatan, pariwisata, ekonomi dan sosial. Kajian ini merekomendasikan beberapa kegiatan yang perlu dilakukan untuk mencegah penyebaran virus. Kegiatan tersebut mencakup deteksi awal, pengawasan pergerakan orang, pengawasan lingkungan, pemeriksaan sarana transportasi dan distribusi barang kebutuhan masyarakat.(Sukmana, Aminuddin and Nopriyanto, 2020). Telaumbanua dalam makalahnya menganalisis berbagai peraturan perundang-undangan yang diperlukan dalam mempercepat penanganan dampak Covid-19 merupakan penyakit menular yang berpotensi Penelitian ini merekomendasikan dua jenis peraturan perundang-undangan yang penting dan mendesak yaitu Peraturan Pemerintah yang wajib dibentuk dalam rangka melakukan tindakan penanggulangan dan pencegahan ancaman penyakit yang mudah menular seperti Covid- 19 serta Peraturan Menteri Kesehatan terkait yang wajib dibentuk dalam rangka

mengantisipasi ancaman Covid-19. (Telaumbanua, 2020). Pandemi Covid-19 juga mempengaruhi sektor pendidikan, yang mana berdasarkan penelitian Wajdi dkk terhadap kebijakan *school from home* (SFH) dan *work from home* (WFH), hal ini dapat meminimalisir penyebaran virus. Kebijakan yang telah ditempuh pemerintah di antaranya pembelajaran daring untuk siswa sekolah dan mahasiswa, penghapusan Ujian Nasional (UN) dan penundaan UTBK SBMPTN 2020. (Wajdi et al., 2020)

Penerapan Teknologi Informasi dan Komunikasi (TIK) juga menjadi salah satu topik spesifik yang dilakukan. Penelitian Zein mengembangkan algoritma kecerdasan buatan untuk menganalisis gambar X-ray dari CT Scan. Dalam penelitian ini, akurasi identifikasi perbedaan gambar antara pneumonia yang terinfeksi virus coronavirus yang dicurigai, sedikit dicurigai, dan pneumonia yang tidak terinfeksi virus corona diklaim cukup tinggi dengan waktu yang singkat. (Zein, 2020). Penelitian studi awal dilakukan oleh Swastika dkk yang mengamati kemungkinan penggunaan citra CT dari organ *thorax* untuk mendeteksi infeksi COVID-19. Data citra CT dimodelkan menggunakan algoritma *deep learning* dengan pengaturan CNN berarsitektur VGG16 dan optimizer SGD dan Adam.(Swastika et al., 2020). Tosepu dkk melakukan kajian dan analisis terhadap korelasi antara cuaca dan pandemi Covid-19 di kota Jakarta. Berdasarkan analisis statistik, suhu di Jakarta mempengaruhi secara signifikan penyebaran virus.(Tosepu et al., 2020). Selanjutnya Nuraini, Khairudin dan Apri menyampaikan hasil kajiannya yang mengusulkan model sederhana untuk memprediksi pandemi Covid-19 di Indonesia. Pemodelan didasarkan atas kurva Richard yang merupakan modifikasi dari persamaan logistik.(Nuraini, Khairudin and Apri, 2020). Hariyadi menggunakan pendekatan Forensik Digital untuk mendapat informasi berupa riwayat berpergian berbasis ponsel cerdas. Pada penelitian ini informasi yang disajikan adalah hasil analisis riwayat berpergian berupa lokasi-lokasi yang pernah dikunjungi oleh penderita.(Hariyadi, 2020). Asyary dan Veruswati melakukan studi tentang korelasi berjemur sinar matahari dengan status Covid-19 di kota Jakarta, Indonesia. Hasil analisis statistik menunjukkan adanya korelasi positif antara aktivitas berjemur sinar matahari dengan tingkat kesembuhan.(Asyary and Veruswati, 2020)

Terkait dengan data dan informasi kegiatan litbang Covid-19, beberapa peneliti telah berupaya untuk mengkaji berbagai data informasi yang terhimpun dalam pangkalan data Covid-19. Penelitian Irawan menguraikan telaah bibliometrika terhadap literatur tentang COVID-19/NCOV/Virus Corona yang dihimpun dalam pangkalan data komersial dan non komersial. Hasil penelitian ini berupaya menggambarkan profil publikasi ilmiah Covid-19 di dunia dengan cara penilaian singkat berdasarkan prinsip-prinsip akses terbuka (*open*

access) dalam diseminasi informasi.(Irawan et al., 2020). Data dan informasi Covid-19 yang terhimpun di berbagai pangkalan data ilmiah dapat digunakan untuk melakukan analisis dan pemetaan topik penelitian berbasis teks. Salah satu metode yang umum digunakan dalam kajian seperti ini adalah klastering berbasis teks. Kata-kata yang dikandung dalam dokumen penelitian diproses dan dipetakan menggunakan algoritma klastering dalam *machine learning*. Penelitian Hardi, Kusuma dan Basuki melakukan analisis dan pemetaan dokumen siaran pers kepunyaan Kedutaan Besar Australia di Jakarta. Dengan menggunakan algoritma *k-means* dan mengatur parameter yang optimal, proses klastering dapat dilakukan dan menghasilkan kluster-kluster yang terpisah secara baik. (Hardi, Kusuma and Basuki, 2019). Hal yang sama dapat diterapkan pada dataset dokumen data dan informasi penelitian Covid-19 dengan menerapkan algoritma klastering *k-means* berbasis teks.

Terkait dengan klastering berbasis graf, Liu dan Barahona melakukan penelitian tentang penggabungan data graf dengan Stabilitas Markov, yaitu framework deteksi komunitas multiskala.(Liu and Barahona, 2020). Sedangkan Jin dan Bai melakukan perbandingan penerapan algoritma klastering hirarki dan *k-means* terhadap data graf hasil *text mining* dan *word co-occurrence*.(Jin and Bai, 2016). Berbeda dengan penelitian klastering dokumen berbasis teks yang telah dilakukan oleh para peneliti lain, dalam penelitian kami, dokumen data dan informasi penelitian Covid-19 dari dataset Kemenristek/BRIN dimodelkan sebagai graf bipartit lalu dilakukan proses klastering. Proses ini menerapkan dua algoritma dan membandingkan hasilnya mana yang memberikan pemetaan kluster-kluster topik penelitian Covid-19 di Indonesia yang lebih baik. Algoritma yang digunakan adalah kernel *k-means* dan *k-means* itu sendiri. Sedangkan pengujian kinerja algoritma klastering dilakukan dengan menerapkan validasi internal indeks Dunn. Sebagai ilustrasi, Gambar 2 di bawah ini menggambarkan model graf dimaksud.

Gambar. 2 Model Graf Bipartit untuk Jaringan Dokumen Istilah

Hasil kajian ini diharapkan dapat memberikan tambahan informasi yang mendukung program pemerintah dalam mempercepat penanganan Covid-19 di Indonesia.

2. METODE PENELITIAN

Analisis didasarkan atas *text mining* dengan hasil yang diinginkan berupa klastering topik penelitian Covid-19. Metode penelitian yang digunakan dalam penelitian ini terinspirasi dari penelitian (Hardi, Kusuma and Basuki, 2019) tentang klastering berbasis teks, penelitian (Nugroho et al., 2019) tentang penerapan kernef graf untuk klasifikasi dan penelitian (Brock et al., 2008) tentang validasi internal untuk klastering. Alur penelitian ini dapat dilihat pada Gambar 3 berikut ini.

Gambar. 3 Alur Penelitian

Dalam penelitian ini dataset dikumpulkan dari situs *Database of Researches, Publications, and Innovations on Coronavirus Disease (Covid-19)* <http://sinta.ristekbrin.go.id/covid/> dengan metode *web scrapping*. Dataset dokumen penelitian Covid-19 yang berhasil dihimpun hingga 6 Agustus 2020 sebanyak 351 judul penelitian. Gambar 4 di bawah ini merupakan ilustrasi dari sebagian dataset yang berhasil dihimpun.

Kemudian dilakukan pra-pemrosesan dengan menghilangkan data N/A sebanyak 75 data dan duplikasi sebanyak 152 data sehingga tersisa 199 data yang bisa digunakan untuk analisis dan pemetaan. Terjadinya data duplikasi kemungkinan disebabkan oleh entri ganda, peneliti dapat mengisi sendiri data penelitian yang dilakukannya dan akan langsung terekam dan ditampilkan di pangkalan data Covid-19. Tahapan selanjutnya adalah menghilangkan *stopwords* dari daftar kata dalam dataset. Kemudian mengambil 500 kata unik sebagai kamus untuk proses klastering dan dibuat sebagai vektor kata. Lalu membuat corpus dari ringkasan dokumen penelitian Covid-19 untuk dibuat matriks istilah dokumen dan model graf bipartit. Selanjutnya dilakukan visualisasi dan analisis karakteristik graf yang terbentuk

Untuk menentukan nilai optimal *k* yang mencerminkan jumlah kluster, dilakukan penilaian menggunakan teknik *plot elbow* yang mana proses inisiasi klastering dilakukan iterasi hingga 100 nilai *k*.

	institusi	judul	nama_peneliti
1	ristekbrin	5 Senyawa Untuk Obat Covid-19	SOETJIPTO
2	ristekbrin	Bahan untuk Hancurkan Material Virus Corona	IKA PUSPITA SARI
3	ristekbrin	Potensi Jeruk Untuk Tangkal Corona	ROHMAD YUDI UTOMO
4	ristekbrin	Zika virus seropositivity in 1-4-year-old children, Ind...	AMIN SUBANDRIO W KUSUMO
5	ristekbrin	Impact of National Disaster Covid-19, Indonesia To w...	ENDANG NARYONO
6	ristekbrin	Pembuatan Hand Sanitizer dan Antiseptik Berbahan...	RINTO M. NUR
7	ristekbrin	Corona Hadis	WAHYUDIN DARMALAKSANA
8	ristekbrin	Analisis Pembelajaran Online Masa WFH Pandemic C...	WAHYUDIN DARMALAKSANA
9	ristekbrin	Analisis Pembelajaran Online Masa WFH Pandemic C...	WAHYUDIN DARMALAKSANA
10	ristekbrin	Analisis Pembelajaran Online Masa WFH Pandemic C...	WAHYUDIN DARMALAKSANA
11	ristekbrin	Upaya Pencegahan Covid-19 Melalui Tradisi Pipiruna...	SARIP
12	ristekbrin	UPAYA PENCEGAHAN COVID-19 DI DESA MELALUI TR...	SARIP
13	ristekbrin	Analisis Dampak Pandemic Covid-19 pada penylene...	SUSILO
14	ristekbrin	Analisis Dampak Pandemic Covid-19 pada penylene...	SUSILO
15	ristekbrin	Kualitas Tata Kelola [Governance] Penanganan. Pen...	ANDRI MEIRIKI
16	ristekbrin	Analisis Dampak Pandemic Covid-19 pada Kegiatan ...	SUSILO
17	ristekbrin	MODEL KOORDINASI PADA MANAJEMEN RANTAI PAS...	Visit Original
18	ristekbrin	Model Koordinasi Pada Manajemen Rantai Pasok Pan...	TOMY PERDANA
19	ristekbrin	PEMANFAATAN APLIKASI GOOGLE SEARCH UNTUK K...	LAILI RAHMATUL ILMI
20	ristekbrin	Pemanfaatan Aplikasi Google Search untuk Kasus C...	LAILI RAHMATUL ILMI
21	ristekbrin	Psychopragmatic Study: A Pattern of Spread Diseas...	WAHYUDI RAHMAT
22	ristekbrin	Pengembangan Model Pembelajaran Blended Learni...	ABDUL RAHMAT
23	ristekbrin	Modeling of COVID-19 Infection per 1 Million Popul...	Visit Original
24	ristekbrin	Imposing lockdown in Indonesia: How long will it tak...	SWANDARI PARAMITA
25	ristekbrin	Imposing lockdown in Indonesia: How long will it tak...	SWANDARI PARAMITA

Gambar. 4 Tampilan Layar Dataset Hasil Web Scrapping

Sesuai dengan kaidah plot *elbow*, posisi kurva *elbow* (mulai melengkung) ada pada nilai $k = 10$, sehingga klastering selanjutnya dilakukan pada nilai ini dan menunjukkan jumlah klusters yang diperoleh dari dataset penelitian Covid-19. Gambar 5 berikut ini menampilkan plot nilai optimal k .

Gambar. 5 Plot Elbow Nilai Optimal k

Setelah itu dilanjutkan tahapan klastering yang dilakukan dengan algoritma kernel *k-means* dan *k-means* dengan nilai $k = 10$. Setelah mendapatkan hasil klastering, tahapan berikutnya adalah menguji kinerja algoritma klastering dengan menggunakan validasi internal indeks Dunn. Indeks Dunn digunakan karena dalam dataset tidak tersedia informasi awal mengenai label atau nama dari masing-masing klaster.

Indeks Dunn didefinisikan sebagai perbandingan antara jarak terkecil antar klaster dengan jarak terbesar dalam klaster. Indeks tersebut dirumuskan sebagai berikut:

$$Dunn = \frac{dist_{inter-min}}{dist_{intra-max}}$$

di mana $dist_{inter-min}$ menyatakan jarak terkecil antara dua titik dari dua klaster yang berbeda, sedangkan $dist_{intra-max}$ menyatakan jarak terbesar antara dua titik dari klaster yang sama. Indeks Dunn dapat mengambil nilai dalam rentang 0 dan ∞ dan seharusnya dimaksimalkan. Nilai indeks Dunn yang lebih besar membuat pasangan yang kuat antara titik-titik data dan keanggotaan klaster. (Brock et al., 2008)

3. HASIL DAN PEMBAHASAN

Pemodelan dataset penelitian Covid-19 sebagai graf bipartit didasari atas adanya korelasi antara kandungan kata-kata unik dalam setiap dokumen dengan kemiripan antar dokumen. Kata-kata yang sama yang muncul dalam beberapa dokumen yang berbeda menandakan adanya kemiripan dari dokumen-dokumen tersebut. Kemiripan ini bisa dijadikan tumpuan untuk melakukan analisis dan pemetaan melalui penerapan algoritma klastering yang tepat. Gambar 6 berikut ini menampilkan model graf bipartit dari dataset penelitian ini.

Gambar. 6 Model Graf Bipartit Dataset Penelitian Covid-19

Untuk melakukan proses klastering, terlebih dahulu perlu menentukan cut-off pada graf. Penentuan cut-off ini berdasarkan grafik distribusi densitas yang menggambarkan pengambilan bagian graf yang paling sesuai untuk diterapkannya algoritma klastering. Dari Gambar 7 berikut ini menampilkan densitas dari model graf bipartit yang terbentuk.

Kandungan kata-kata atau istilah unik dalam suatu dataset dapat divisualisasikan sebagai awan kata (*wordclouds*). Menurut Heimerl dkk, penggunaan *wordclouds* dalam berbagai konteks bertujuan untuk menyediakan tinjauan umum dengan mengambil teks berdasarkan frekuensinya. (Heimerl et al., 2014). Pada Gambar 8 di bawah ini, terlihat beberapa istilah penting muncul dengan frekuensi terbanyak. Perbedaan ukuran huruf menunjukkan seberapa banyak istilah tersebut ada dalam dataset, sedangkan perbedaan warna menunjukkan data awal klastering yang mana istilah-istilah dengan warna yang sama kemungkinan besar berada pada klaster yang sama.

Gambar. 7 Distribusi Densitas Model Graf Bipartit

Gambar. 8 Visualisasi Wordcloud Kata-kata Unik Dataset Penelitian Covid-19

Sebagai konfirmasi dari visualisasi wordclouds di atas, Gambar 9 berikut ini menampilkan 20 kata atau istilah unik dengan frekuensi terbanyak.

Gambar. 9 Grafik Frekuensi 20 Kata/Istilah Unik

Sebagai langkah lanjutan untuk melakukan klastering, corpus istilah unik dibangun dan dikonversi menjadi matriks dokumen istilah (*document term matrix*). Setelah itu dilakukan prosedur inspeksi untuk memastikan bahwa matriks dokumen istilah yang dihasilkan bisa digunakan untuk klastering. Hasil inspeksi menunjukkan ketersebaran (*sparsity*) sebesar 95% yang bermakna dataset ini cukup bagus dan dapat digunakan sebagai dasar klastering. Gambar 10 berikut ini menampilkan hasil inspeksi tersebut.

```
> inspect(corpus.dtm)
<<DocumentTermMatrix (documents: 199, terms: 514)>>
Non-/sparse entries: 5118/97168
sparsity: 95%
Maximal term length: 16
weighting: term frequency (tf)
Sample terms:
```

docs	corona	data	hasil	indonesia	online	pandemi	pasien	pembelajaran	penelitian
12	0	2	2	3	0	8	3	0	8
13	0	2	2	3	0	8	3	0	8
132	1	2	0	2	2	0	1	3	3
181	1	0	1	4	0	0	2	0	3
182	1	0	1	4	0	0	2	0	3
65	3	3	1	4	0	0	2	0	0
67	3	3	1	4	0	0	2	0	0
68	3	3	1	4	0	0	2	0	0
69	3	3	1	4	0	0	2	0	0
70	3	3	1	4	0	0	2	0	0

docs	virus
12	0
13	0
132	1
181	7
182	7
65	5
67	5
68	5
69	5
70	5

Gambar. 10 Hasil Inspeksi Matriks Dokumen Istilah

Dalam penemuan pengetahuan (*knowledge discovery*), seringkali ditemukan kumpulan informasi melimpah baik dalam bentuk tekstual maupun struktur graf. Menurut Mei dkk, model statistik topik dapat digunakan secara bersama dengan struktur graf untuk mendapatkan pemetaan topik dokumen penelitian. (Mei et al., 2008). Dataset penelitian Covid-19 dari Kemenristek/BRIN dapat dipetakan menjadi setidaknya 10 topik berdasarkan algoritma klastering yang digunakan. Gambar 11 berikut ini menunjukkan adanya 10 klaster topik penelitian dan setiap klaster diwakili oleh 5 istilah utama yang mencerminkan kata kunci dari masing-masing klaster.

Gambar. 11 Graf Pemetaan Topik Penelitian Covid-19

Dari graf topik penelitian, kita bisa mengetahui bahwa terdapat tumpang tindik topik yang dicerminkan dari adanya kesamaan istilah yang muncul dalam beberapa topik penelitian. Gambar 12

berikut ini menunjukkan adanya tumpah tindih tersebut.

Gambar. 12 Graf Tumpang Tindih Istilah Unik dalam Peta Topik Penelitian

Sebagai langkah validasi terhadap proses klastering, maka dalam penelitian ini dilakukan validasi internal menggunakan indeks Dunn. Validasi internal dipilih karena dalam dataset tidak terdapat informasi awal masing-masing dokumen telah terkelompokkan ke klaster tertentu sehingga yang paling mungkin untuk uji validitas adalah menggunakan validasi internal. Salah satu metode uji validitas internal untuk klastering adalah penghitungan indeks Dunn. Tabel 1 berikut ini menampilkan hasil perhitungan indeks Dunn terhadap penerapan algoritma kernel *k-means* dan *k-means*.

Tabel. 1 Perbandingan Indeks Dunn Algoritma *k-means* dan kernel *k-means*

dunn.kmeans	dunn.kkmeans
0.2069141	0.3779645

Dari indeks Dunn di atas dapat disimpulkan bahwa kedua algoritma menghasilkan hasil klastering yang baik, paling sedikit sepasang klaster terpisah dengan baik dan/atau seluruh klaster terpisah dengan baik. Dalam hal penelitian ini, 10 klaster terpisah dengan baik dan kompak. Namun algoritma kernel *k-means* menunjukkan hasil yang sedikit lebih daripada *k-means*. Algoritma kernel *k-means* memberikan hasil yang lebih unggul disebabkan oleh kemampuan algoritma ini menangani dataset yang bersifat non-linier yang mana dalam penelitian ini dataset penelitian Covid-19 dari Ristekbrin dimodelkan sebagai graf bipartit sebagaimana telah diungkapkan oleh Chitta dkk.(Chitta et al., 2011)

4. KESIMPULAN

Data dan informasi penelitian Covid-19 yang disediakan untuk public oleh Kemenristek/BRIN telah berhasil dianalisis dan dipetakan dalam penelitian ini. Algoritma klastering berbasis teks yaitu *k-means* dan algoritma berbasis kernel graf yaitu kernel *k-means* telah digunakan dan

menghasilkan peta 10 klaster topik penelitian Covid-19 di Indonesia. Perbandingan kinerja kedua algoritma telah diuji menggunakan metode validasi internal indeks Dunn. Dari perhitungan indeks Dunn, kedua algoritma terbukti berhasil baik untuk penerapan klastering dataset Covid-19. Namun, kernel *k-means* memberikan hasil klastering yang sedikit lebih baik dibandingkan dengan *k-means*. Hasil analisis dan pemetaan ini memberikan tambahan informasi bagi pemerintah selaku pengambil kebijakan terutama terkait kebijakan litbang Covid-19 agar memperhatikan 10 klaster yang telah ditemukan dalam penelitian ini. Selain itu, bagi para akademisi, dengan adanya peta 10 klaster ini, diharapkan mampu melihat peluang litbang yang lebih menjanjikan dan menghasilkan luaran yang mampu membantu mempercepat penanganan Covid-19 di Indonesia.

Penelitian ini masih terbatas pada dataset penelitian yang diunggah langsung oleh para peneliti. Pengembangan penelitian ini diharapkan bisa mencakup dataset lainnya yaitu dataset publikasi ilmiah dan dataset data ilmiah primer yang juga dihimpun oleh Kemenristek/BRIN. Selain itu, penggunaan kernel graf perlu diperluas dengan kernel yang lebih banyak sehingga menghasilkan peta yang lebih komprehensif.

DAFTAR PUSTAKA

- ASYARY, A. & VERUSWATI, M., 2020. Sunlight exposure increased Covid-19 recovery rates: A study in the central pandemic area of Indonesia. *Science of The Total Environment*, [online] 729, p.139016. Available at: <<https://doi.org/10.1016/j.scitotenv.2020.139016>>.
- BROCK, G., PIHUR, V., DATTA, S. & DATTA, S., 2008. cValid: An R Package for Cluster Validation. *Journal of Statistical Software*, [online] 25(4), pp.1–32. Available at: <<http://www.jstatsoft.org/v25/i04/>>.
- CHITTA, R., JIN, R., HAVENS, T.C. & JAIN, A.K., 2011. Approximate Kernel K-Means: Solution to Large Scale Kernel Clustering. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*. [online] New York, NY, USA: Association for Computing Machinery.pp.895–903. Available at: <<https://doi.org/10.1145/2020408.2020558>>.
- HARDI, W., KUSUMA, W.A. & BASUKI, S., 2019. Clustering topic groups of documents using K-Means algorithm: Australian Embassy Jakarta media releases 2006-2016. *Berkala Ilmu Perpustakaan dan Informasi*, 15(2), p.226.
- HARIYADI, D., 2020. Investigasi Dini Contact

- Tracing Pasien Menggunakan Pendekatan Standarisasi Forensik Digital. *Jurnal Repositor UMM*, 2(5), pp.583–590.
- HEIMERL, F., LOHMANN, S., LANGE, S. & ERTL, T., 2014. Word Cloud Explorer: Text Analytics Based on Word Clouds. In: *2014 47th Hawaii International Conference on System Sciences*. pp.1833–1842.
- IRAWAN, D.E., PANGARSO, A., RIDLO, I.A. & FUAD, A., 2020. Telaah bibliometrik pola penyebaran pengetahuan tentang COVID-19 di dunia. *Jurnal Matematika dan Sains Institut Teknologi Bandung*, [online] (April), pp.1–15. Available at: <https://figshare.com/articles/The_rapid_publications_on_COVID_An_open_science_perspective/12084339/3>.
- JIN, C. AND BAI, Q., 2016. Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence. In: *2016 International Conference on Information System and Artificial Intelligence (ISAI)*. pp.497–502.
- LIU, Z. & BARAHONA, M., 2020. Graph-based data clustering via multiscale community detection. *Applied Network Science*, [online] 5(1), p.3. Available at: <<https://doi.org/10.1007/s41109-019-0248-7>>.
- MEI, Q., CAI, D., ZHANG, D. & ZHAI, C., 2008. Topic Modeling with Network Regularization. In: *Proceedings of the 17th International Conference on World Wide Web, WWW '08*. [online] New York, NY, USA: Association for Computing Machinery. pp.101–110. Available at: <<https://doi.org/10.1145/1367497.1367512>>.
- NUGROHO, B., ARITSUGI, M., OTACHI, Y. & MANABE, Y., 2019. Combined graph kernels for automatic patent classification: A hybrid approach. *World Patent Information*, [online] 57, pp.18–24. Available at: <<http://www.sciencedirect.com/science/article/pii/S0172219018300577>>.
- NURAINI, N., KHAIRUDIN, K. & APRI, M., 2020. Modeling Simulation of COVID-19 in Indonesia based on Early Endemic Data. *Communication in Biomathematical Sciences*, 3(1), pp.1–8.
- SUKMANA, M., AMINUDDIN, M. & NOPRIYANTO, D., 2020. Indonesian government response in COVID-19 disaster prevention. *East African Scholars Journal of Medical Sciences*, [online] 3(3), pp.81–6. Available at: <<https://www.easpublisher.com/easjms>>.
- Swastika, W., Studi, P., Informatika, T. & Korespondensi, P., 2020. Studi Awal Deteksi Covid-19 Menggunakan Citra Ct Berbasis Deep Preliminary Study of Covid-19 Detection Using Ct Image Based on. 7(3), pp.629–634.
- TELAUMBANUA, D., 2020. Urgensi Pembentukan Aturan Terkait Pencegahan Covid-19 di Indonesia. *QALAMUNA: Jurnal Pendidikan, Sosial, dan Agama*, 12(01), pp.59–70.
- TOSEPU, R., GUNAWAN, J., EFFENDY, D.S., AHMAD, L.O.A.I., Lestari, H., Bahar, H. and Asfian, P., 2020. Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, 725.
- WAJDI, M.B.N., IWAN KUSWANDI, UMAR AL FARUQ, ZULHIJRA, Z., KHAIRUDIN, K. AND KHOIRIYAH, K., 2020. Education Policy Overcome Coronavirus, A Study of Indonesians. *EDUTEC: Journal of Education And Technology*, 3(2), pp.96–106.
- ZEIN, A., 2020. Pendeteksian Vlrus Corona Dalam Gambar X-Ray Menggunakan Algoritma Artifical Intelligence Dengan Deep Learning Python. *jurnal Teknologi Informasi ESIT*, XV(01), pp.19–23.

Halaman ini sengaja dikosongkan