Klasterisasi Menggunakan Algoritma *K-Means* dan *Elbow* pada Opini Masyarakat Tentang Kebijakan Sekolah Luring Tahun 2022

Rahmawan Bagus Trianto[1], Agus Susilo Nugroho[2], Eko Supriyadi[3]
[1, 2, 3] Universitas An Nuur Jl. Gajah Mada No. 7 Majenang, Kuripan, Kec. Purwodadi, Kab. Grobogan, Jawa Tengah, Indonesia
Email: rahmawanbagust@gmail.com[1], nugro333@gmail.com[2], ekalaya56@gmail.com[3]

Abstract - The covid-19 pandemic that swept across the globe had adverse effects in many areas. One of the most affected areas is education in Indonesia. The online learning model became the only option at the time, which had a negative impact on the quality of education in Indonesia. As time went on, conditions are getting better, but there was still a threat of covid-19. In early 2022 governments began to adopt face-to-face or offline learning that attracted opinions on social media. The opinions that are widely written on social media need to be prepared because they could be input to the government. Clustering using the *k-means* algorithm with the *elbow* method as its optimizer in determining the best cluster number is one of the opinions processing options on social media for measuring and accounting. Data is treated with two approaches: with and without *stemming*. Applying the *elbow* method to the *k-means* algorithm produces a performance of the clustering model with a DBI value of 0.003 with 4 clusters, and a value of SSE 0.331, for data without *stemming*. On data with treatment using *stemming*, it has 3 cluster numbers with a value of DBI at 0.003 and SSE at 0426.

Keywords – *K-Means*, *elbow*, *stemming*, social media, opinion.

Intisari – Pandemi covid-19 yang melanda hampir seluruh belahan dunia berdampak negatif di berbagai bidang. Salah satu bidang yang terdampak adalah pendidikan di Indonesia. Model pembelajaran daring menjadi satu-satunya pilihan pada waktu itu, sehingga menimbulkan dampak buruk bagi kualitas pendidikan di Indonesia. Seiring berjalannya waktu, kondisi semakin membaik, namun masih ada ancaman bahaya covid-19. Pada awal tahun 2022 pemerintah mulai menerapkan pembelajaran tatap muka atau luring yang mengundang berbagai opini di media sosial. Opini yang banyak tertulis di media sosial perlu diolah karena memiliki pendapat yang bisa menjadi masukan bagi pemerintah. Klasterisasi menggunakan algoritma *k-means* dengan metode *elbow* sebagai optimasinya dalam menentukan jumlah klaster terbaik menjadi salah satu pilihan pengolahan opini di sosial media agar dapat diukur dan dipertanggungjawabkan. Data diperlakukan dengan dua pendekatan, yaitu diolah dengan menggunakan *stemming* dan tanpa *stemming*. Dengan menerapkan metode *elbow* pada algoritma *k-means* menghasilkan performa model klasterisasi dengan nilai DBI 0.003 dengan klaster sebanyak 4 buah, dengan nilai SSE sebesar 0.331 untuk data tanpa *stemming*. Pada data dengan perlakuan dengan menerapkan *stemming* memiliki jumlah klaster sebanyak 3 buah dengan nilai DBI sebesar 0.003 dan SSE sebesar 0.426.

Kata Kunci – *K-Means*, *elbow*, *stemming*, sosial media, opini.

I. PENDAHULUAN

Pandemi covid-19 menerjang hampir seluruh penjuru dunia yang berdampak negatif di berbagai bidang, salah satunya adalah pendidikan di Indonesia[1]. Pada awal pandemi covid-19 pilihan model pembelajaran dilakukan secara daring untuk menekan pertumbuhan kasus. Seiring dengan berjalannya waktu, kondisi pandemi semakin membaik meskipun tetap terdapat ancaman kesehatan, maka dilakukan model pembelajaran secara *hybrid*. Hal ini dilakukan

untuk menjaga kualitas pendidikan di Indonesia dan juga dapat memperkecil resiko buruk, yang juga disesuaikan dengan situasi dan kondisi di wilayah tersebut [1]. Namun demikian, kualitas pendidikan di masa pandemi tetap memberikan dampak buruk seperti penurunan waktu pembelajaran yang hampir satu tahun jika dibandingkan sebelum pandemi [2].

Pemerintah dalam menangani pandemi covid-19 ini terutama di bidang pendidikan, mengeluarkan kebijakan pendidikan tatap muka yang dimulai pada tahun ajaran 2021-2022 semester ganjil. Pelaksanaan dari kebijakan ini tentunya dilakukan dengan kontrol yang ketat, seperti penerapan protokol kesehatan yang sesuai anjuran[3]. Masyarakat yang mengetahui kebijakan ini ramai meluapkan pendapat dan opininya di sosial media masing-masing. *Twitter* menjadi satu dari banyak sosial media yang paling aktif pergerakan datanya, di mana berisi tulisan singkat namun dengan jumlah yang sangat banyak [4][5].

Masyarakat yang merespon kebijakan tersebut meluapkan opininya di media sosial *twitter* sebenarnya tidak hanya sekedar tulisan tanpa arti, akan tetapi bisa berisi opini dan pendapat tentang harapan dan keinginan terkait pandemi covid-19 ini. Pandangan, harapan, keinginan dari masyarakat satu memiliki keterkaitan dan kemiripan dengan masyarakat yang lain, yang berarti adanya masukan yang bagus bagi institusi khususnya institusi pendidikan bahkan pemerintah [6]. Dari sana dapat diambil langkah untuk dilakukan pengolahan data dari sosial media seperti *twitter* untuk kemudian dianalisis informasi yang terkandung dari curahan hati masyarakat. Dengan demikian harapan, keinginan dan pandangan masyarakat tersebut dapat tersampaikan dengan baik untuk institusi terkait dan juga pemerintah [7]. Dengan adanya masukan yang sudah diberikan masyarakat melalui media sosial dapat memberikan opsi lain bagi pemegang kepentingan dalam menentukan kebijakan yang tepat bagi dunia pendidikan Indonesia.

Metode pengelompokan data atau juga dikenal dengan istilah klastering telah dipakai di banyak penelitian. Klastering dilakukan dengan melihat tingkat kemiripan dari setiap data, yang kemudian dikelompokkan dalam kelompok yang sama, sehingga data antar kelompok memiliki tingkat kemiripan yang kecil dan data di kelompok yang sama memiliki tingkat kemiripan yang tinggi. Dengan demikian, pengelompokan data dengan metode klastering tidak memerlukan adanya label di dataset yang akan dipakai. Metode yang banyak dipakai untuk melakukan klastering data teks antara lain *k-means* [8], *k-medoids* [9], serta *Latent Dirichlet Allocation* (LDA) [10]. Dari beberapa algoritma tersebut, algoritma *k-means* memiliki performa yang paling baik [11]. Akan tetapi, algoritma *k-means* memiliki kekurangan dalam menentukan nilai *k* dan juga *centroid* yang mengakibatkan performanya tidak optimal [12][13]. Untuk dapat memperbaiki performa algoritma *k-means* maka dapat dipakai metode *elbow*, yaitu metode untuk menentukan nilai *k* yang optimal [14].

Adapun tujuan dari penelitian ini adalah menggabungkan metode *elbow* dengan algoritma *k-means* dengan harapan dapat meningkatkan performa model klastering. Dengan diterapkannya metode *elbow* untuk algoritma *k-means* yang dapat mengoptimalkan hasil klastering, semakin optimal juga pengelompokan data opini masyarakat terhadap kebijakan sekolah tatap muka di masa pandemi, sehingga dapat dijadikan salah satu bahan masukan dalam membuat kebijakan ke depannya.

II. SIGNIFIKANSI STUDI

A. Studi Literatur

Text mining merupakan salah satu cabang dalam ilmu komputer yang mempelajari tentang ekstraksi dokumen teks secara otomatis, sehingga informasi yang terdapat di dalamnya dapat dimunculkan [15]. Dalam penerapannya, *text mining* banyak dilakukan untuk kebutuhan marketing, perencanaan produksi, pemasaran digital, dengan sumber datanya berupa teks dari media sosial [16]. Pemanfaatan *text mining* dengan menggunakan data dari media sosial

dikarenakan tulisan dan ulasan dari masyarakat memiliki pengaruh di dunia marketing, sehingga dengan mengolah data teks menggunakan teknik *text mining* dapat menghasilkan informasi dan pengetahuan yang bisa dijadikan dasar dalam pengambilan keputusan.

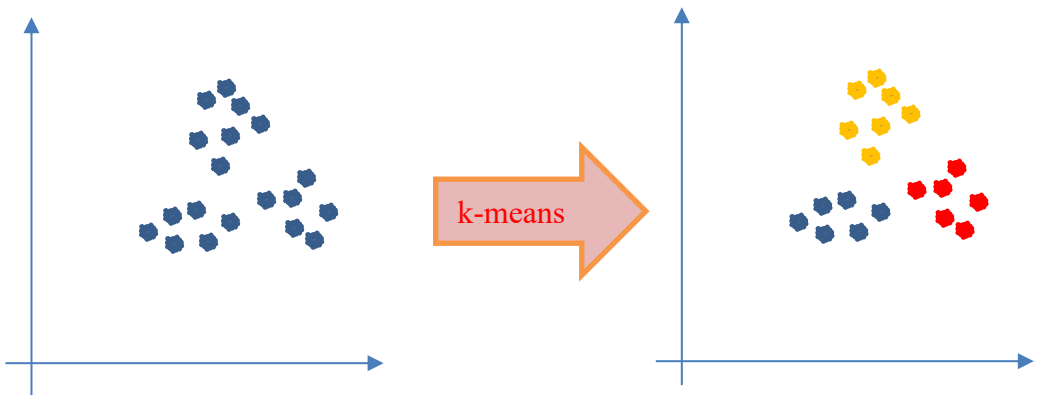
Dalam *text mining* terdapat tahap *preprocessing* yang merupakan tahap yang cukup menentukan terhadap hasil akhir performa model yang terbentuk. Setelah tahap ini dilakukan, perlu dihitung bobot setiap kata atau *term* pada dokumen teks. Pembobotan ini menggunakan *Term-Frequency Inverse-Document-Frequency* atau yang disingkat TF-IDF. TF-IDF ini dapat dihitung menggunakan persamaan 1 dan 2 berikut.

$$W_{i,j} = TF_{i,j} * IDF_j \quad (1)$$

$$IDF_j = log \frac{N}{DF_j} \quad (2)$$

Di mana:
W*i,j* adalah bobot dari kata ke i pada dokumen ke j
TF*i,j* adalah frekuensi kemunculan *term* i pada dokumen j
IDF*j* adalah *inverse document frequency* pada term ke j
N adalah jumlah seluruh dokumen
DF*j* adalah banyaknya dokumen yang mengandung term j
Dimana dokumen di sini merupakan dokumen tweet.

K-means merupakan salah satu algoritma klastering yang mengelompokkan *N* buah data ke dalam *k* buah klaster dengan cara meminimalkan jumlah kuadrat jarak atau *sum of square* antara setiap titik data dan mendekatkan ke titik pusat klaster tersebut [17]. Ilustrasi dari algoritma *k-means* ini dapat dilihat pada gambar 1 di bawah ini. Pada gambar 1 terlihat data awal tersebar tanpa adanya label yang ditandai dengan warna yang sama. Kemudian data tersebut diolah dengan menggunakan *k-means* dibuat tiga buah klaster, yang ditandai data yang berkelompok sesuai dengan tingkat kemiripannya, yang ditandai dengan tiga buah warna yang berbeda.



Gambar 1. Ilustrasi algoritma *k-means*

Algoritma *k-means* dapat dijelaskan sebagai berikut [13]:

a. Menentukan jumlah klaster *k* dan memilih pusat klaster atau *centroid* secara acak.
b. Menghitung jarak setiap data ke pusat klaster menggunakan *Euclidean distance* dengan persamaan berikut:

$$d_{i,j} = \sqrt{\sum_{x=1}^{n} (a_{xi} - b_{xj})^2} \quad (3)$$

Di mana d*i,j* merupakan jarak dari data ke-*i* dengan pusat klaster ke-*j* yang akan dihitung. Banyaknya data dituliskan dengan nilai *n*.

c. Mengelompokkan data ke dalam klaster berdasarkan jarak terpendek, yang ditandai dengan nilai jarak *Euclidean* yang paling kecil.
d. Menghitung ulang pusat klaster yang baru dengan persamaan berikut:

$$j_{i,k} = \frac{1}{n_k}\sum_{j=1}^{m} x_{j,i} \qquad (4)$$

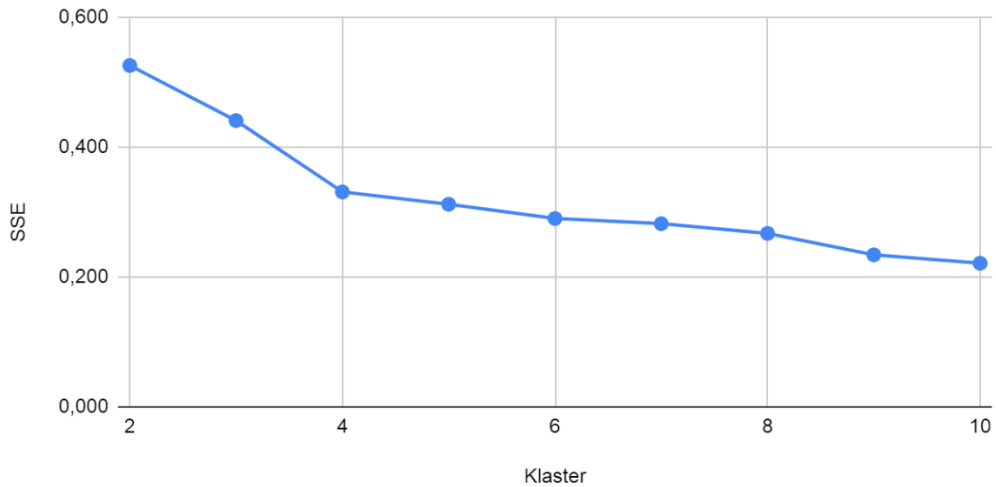
Di mana ji,k merupakan centroid baru yang akan dicari, dengan rata-rata klaster ke-k untuk data ke-i. Kemudian nk merupakan banyaknya data yang menjadi anggota klaster ke-k. Adapun xj,i merupakan frekuensi kemunculan kata ke-j pada dokumen ke-i pada klaster tersebut.

e. Mengulangi langkah 2 sampai 4 hingga data yang berada dalam masing-masing klaster tidak mengalami perubahan.

Metode *elbow* merupakan metode yang dipakai untuk mengoptimalkan performa dari *k-means* dengan cara memilih nilai *k* dengan menghitung SSE atau *sum of square error* [18]. Untuk menghitung nilai SSE dapat menggunakan persamaan berikut:

$$SSE = \sum_{k=1}^{k}\sum_{xi\,\epsilon Sk}\|x_i - C_k\|^2 \qquad (5)$$

Di mana SSE adalah nilai SSE yang akan dicari, *k* adalah banyaknya klaster, *xi* adalah jumlah data dan *Ck* adalah banyaknya cluster *i* pada klaster ke *k*. Algoritma elbow dapat juga dijelaskan cara kerjanya yaitu dengan membandingkan nilai dari k buah klaster dan nilai SSE sehingga membentuk grafik yang menyerupai siku [14]. Grafik yang terbentuk dari perbandingan nilai k dengan SSE dapat dilihat pada gambar 2 berikut.



Gambar 2. Contoh grafik algoritma *elbow* antara nilai SSE dan jumlah klaster

Dengan menggunakan algoritma *elbow* ini maka nilai dari k yang menunjukkan jumlah klaster terbaik adalah sebanyak 4 buah klaster. Hal ini dapat dilihat pada grafik yang memiliki penurunan nilai SSE yang besar, atau sudut yang terbentuk akan cenderung lebih kecil dibandingkan dengan yang lain [19].

Untuk mengukur performa klaster yang terbentuk, perlu dilakukan pengujian. Pada penelitian ini untuk menguji model yang terbentuk menggunakan *Davies Bouldin Index* atau DBI. Untuk menghitung nilai DBI menggunakan konsep memaksimalkan selisih jarak inter

cluster dan meminimalkan selisih jarak intra cluster [14]. Langkah-langkah dalam menghitung nilai DBI dapat mengikuti algoritma berikut.

a. Menghitung nilai *Sum of Square Within* Cluster (SSW) dengan persamaan 6 berikut, dengan mencari matriks kohesi pada sebuah klaster ke-*i*.

$$SSW_i = \frac{1}{m}\sum_{j=1}^{mi} d(xj, ci) \qquad (6)$$

b. Menghitung nilai *Sum of Square Between Cluster* (SSB) dengan persamaan 7 berikut, untuk mencari selisih jarak antar data pada antar klaster.

$$SSB\ i,j = d(ci, cj) \qquad (7)$$

c. Menghitung rasio dengan cara membandingkan nilai klaster ke *i* dengan klaster ke *j* dengan menggunakan persamaan 8 berikut.

$$Ri,j = \frac{SSWi+SSWj}{SSB\ ij} \qquad (8)$$

d. Menghitung nilai DBI setelah mendapatkan nilai rasio dengan persamaan 9 berikut.

$$DBI = \frac{1}{k}\sum_{1=1}^{k} max_{i\neq j}(R_{i,j}) \qquad (9)$$

Penelitian terkait dengan klastering sudah banyak dilakukan oleh para peneliti sebelumnya, baik untuk data teks [20], data citra atau gambar [21], serta data terstruktur [22]. Klasterisasi merupakan teknik pengelompokan data di mana objek yang berada dalam klaster yang sama memiliki tingkat kemiripan yang tinggi, sekaligus objek yang berbeda klaster memiliki tingkat ketidakmiripan yang tinggi [23].

Pada penelitian [12] membahas tentang pemakaian algoritma PSO terhadap algoritma *k-means* untuk klastering daerah endemik penyakit menular di Kota Semarang. Dari penelitian ini menghasilkan nilai DBI sebesar 0.113 pada *k-fold* sebesar 5. Pada penelitian [24] terkait dengan optimasi *k-means* dengan menggunakan *information gain*. Dari penelitian ini didapatkan penurunan nilai DBI menjadi 1.8117 dari nilai awalnya pada angka 2.0290. Penelitian yang lain terkait dengan algoritma *k-means* juga dilakukan pada penelitian [17] di mana dengan menggunakan iterasi sebanyak 100 kali, dapat menurunkan tingkat error dari 15% menjadi hanya 6% saja.

Penelitian [22] membahas tentang pengelompokan desa menggunakan *k-means* dalam rangka penyelenggaraan penanganan bencana banjir. Algoritma *k-means* dioptimasi dengan metode *elbow* dan *silhouette*. Dari percobaan ini menghasilkan bahwa metode *elbow* memiliki performa yang lebih baik dibandingkan dengan *silhouette* pada algoritma *k-means*, yaitu memiliki nilai *variance within* sebesar 2.570625 dan nilai *variance between* sebesar 27.0233857. Selanjutnya penelitian [25] terkait tentang optimasi algoritma *k-means* dengan menggunakan QK-Hidden *algorithm*. Dibandingkan dengan algoritma *k-means* murni, algoritma yang diusulkan pada penelitian ini dapat meningkatkan performa model sampai lebih dari 30%.

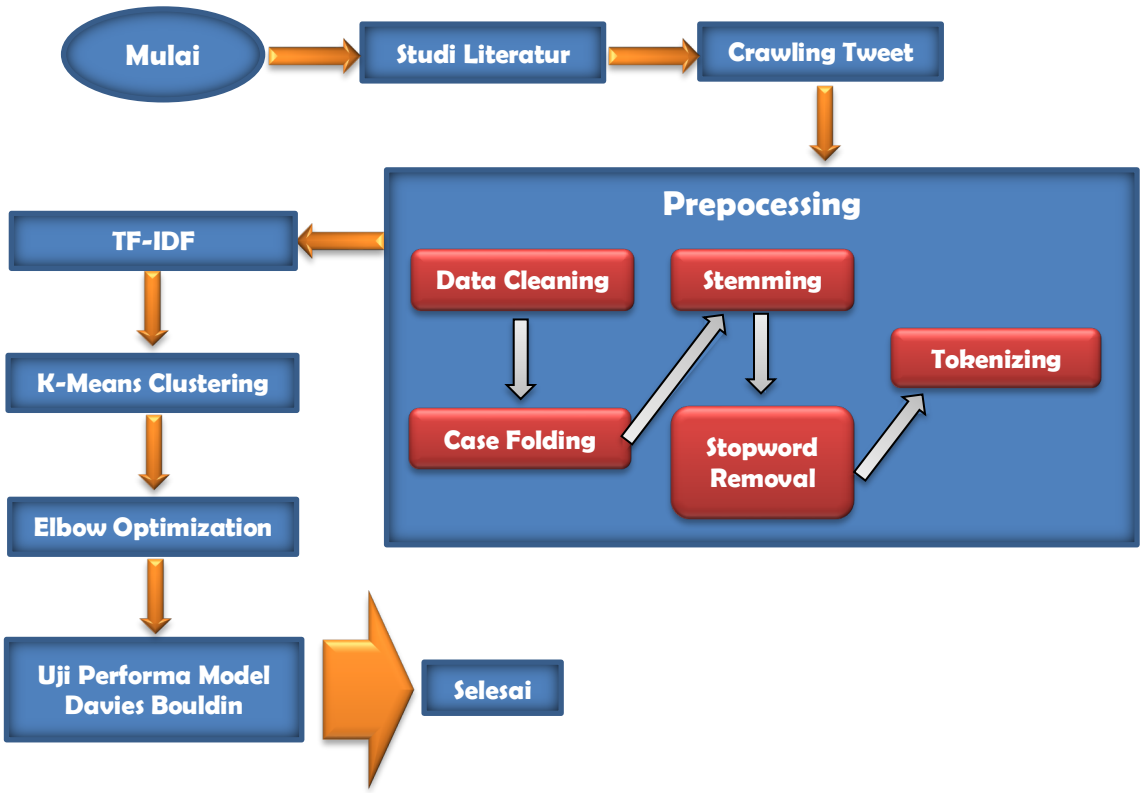
Dari beberapa penelitian yang telah dilakukan peneliti sebelumnya, belum ditemukan klastering dokumen teks dengan menggunakan algoritma *k-means* yang dioptimasi oleh algoritma *elbow*. Jika melihat penelitian [22] di mana algoritma *k-means* dioptimasi dengan metode *elbow* menghasilkan performa yang baik dengan data yang dipakai adalah data terstruktur. Pada penelitian ini akan menggunakan kombinasi algoritma klastering *k-means* dan *elbow* untuk data tidak terstruktur, yaitu data teks opini masyarakat.

B. Dataset

Penelitian ini menggunakan dataset yang bersifat public, yaitu berasal dari media sosial *twitter*. Pengambilan dataset ini menggunakan beberapa kata kunci, yaitu pembelajaran offline, pembelajaran tatap muka, sekolah luring, sekolah offline dan sekolah tatap muka. Setelah didapatkan semua dataset, kemudian diambil sampel sebanyak 1000 buah data. Data yang didapatkan akan diperlakukan dengan dua pendekatan, yaitu data dengan menggunakan *stemming* dan tanpa menggunakan *stemming*. Pendekatan ini diambil karena *stemming* memiliki dua kemungkinan terhadap performa model yang terbentuk, dapat meningkatkan dan juga dapat menurunkan kualitas model.

C. Metode penelitian

Metode penelitian ini dimulai dari studi literatur yang dilanjutkan dengan pengambilan dataset. Dataset diambil dengan cara *crawling* data *twitter* dengan beberapa kata kunci seperti yang telah disebutkan pada poin sebelumnya. Dataset kemudian diambil sampel sebanyak 1000 buah data. Data yang telah diambil diolah terlebih dahulu di tahap *preprocessing*. Tahap *preprocessing* ini dilakukan dengan beberapa proses, dimulai *data cleaning*, *case folding*, *stemming*, *stopword removal* dan *tokenizing*. Khusus untuk proses *stemming*, pada penelitian ini dilakukan juga tanpa menggunakan *stemming*. Dilanjutkan dengan proses pembobotan kata di setiap dokumen menggunakan TF-IDF. Setelah didapatkan bobot dari kata-kata kemudian dilakukan klastering menggunakan *k-means* dan *elbow* yang kemudian dilihat performa model dengan mencari nilai *Davies Bouldin Index* atau DBI. Semakin kecil nilai DBI maka semakin baik pula performa model klasterisasi yang terbentuk. Metode penelitian ini dapat dilihat pada gambar 3 berikut.



Gambar 3. Metode Penelitian

Gambar 3 di atas menunjukkan alur metode penelitian yang dilakukan pada penelitian ini. Selanjutnya dilakukan proses penelitian dan dibahas serta memunculkan hasil seperti yang dijelaskan pada bab selanjutnya.

III. HASIL DAN PEMBAHASAN

Pengumpulan dataset yang telah dilakukan dengan cara *crawling* pada sosial media berbasis teks *twitter* mendapatkan 1000 dataset. Agar dataset ini siap untuk dilakukan pengolahan, perlu perlakuan awal atau tahap *preprocessing* yaitu *data cleaning*, *case folding*, *stemming*, *stopword removal* dan *tokenizing*. Pada penelitian ini pendekatan yang dilakukan pada tahap *preprocessing* ada dua skenario, yaitu menggunakan *stemming* dan tanpa *stemming*. Dua pendekatan ini dilakukan untuk mengetahui seberapa pengaruh penggunaan *stemming* untuk dokumen teks ringkas berbahasa Indonesia. Tahap *preprocessing* ini dapat dilihat pada tabel 1 berikut ini.

TABEL I
TAHAP PREPROCESSING DATASET

Data Awal	Data Cleaning	Stemming	Stopword Removal	Tokenizing
karena sdm indonesia saat ini aja di pendidikan kurang baik, juga murid banyak yang malas. lagi pula, tiap orang memiliki metode pembelajaran yang berbeda-beda, kayak gue yang kebanyakan metode belajarnya harus lama dan tatap muka baru bisa ngerti, ketimbang sebentar	karena sdm indonesia saat ini aja di pendidikan kurang baik, juga murid banyak yang malas. lagi pula, tiap orang memiliki metode pembelajaran yang berbeda-beda, kayak gue yang kebanyakan metode belajarnya harus lama dan tatap muka baru bisa ngerti, ketimbang sebentar	karena sdm indonesia saat ini aja di didik kurang baik juga murid banyak yang malas lagi pula tiap orang milik metode ajar yang beda kayak gue yang banyak metode ajar harus lama dan tatap muka baru bisa ngerti ketimbang sebentar	karena sdm indonesia saat ini aja didik kurang baik juga murid banyak malas lagi pula tiap orang milik metode ajar yang beda kayak gue yang banyak metode ajar harus lama tatap muka baru bisa ngerti ketimbang sebentar	karena sdm indonesia saat ini aja didik kurang baik juga murid banyak malas lagi pula tiap orang milik metode ajar beda kayak gue banyak metode ajar harus lama tatap muka baru bisa ngerti ketimbang sebentar

Setelah tahap *preprocessing* dilakukan, dilanjutkan dengan menghitung pembobotan untuk setiap kata pada dokumen dilakukan menggunakan TF-IDF. Setelah didapatkan bobot dari setiap kata, kemudian dihitung untuk mendapatkan bobot dari setiap dokumen. Pembobotan pada setiap dokumen selanjutnya dilakukan proses klasterisasi. Untuk mendapatkan jumlah klaster yang optimal, penelitian ini menggunakan metode *elbow* di mana didapatkan hasil

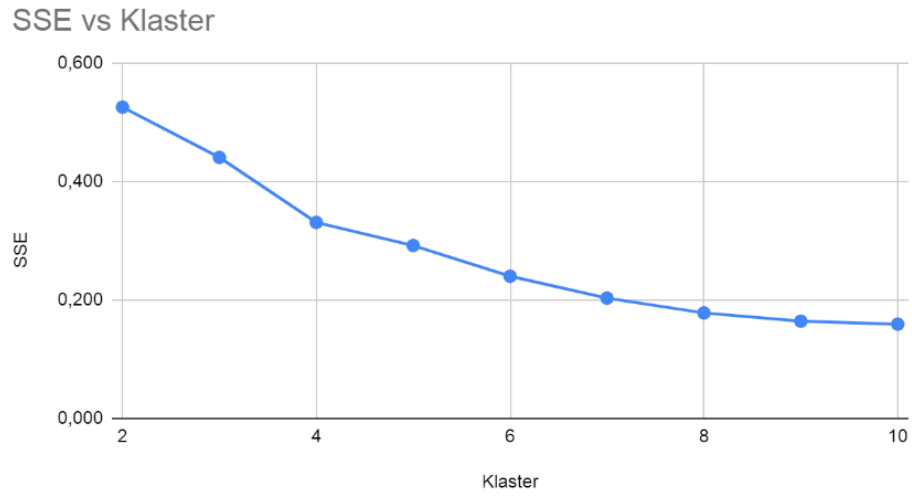
sebagai berikut. Dari pengujian dan percobaan yang telah dilakukan, yaitu menerapkan metode *elbow* pada algoritma *k-means* didapatkan hasil sebagai berikut.

TABEL II
PERBANDINGAN *K-MEANS* DAN METODE *ELBOW* PADA DATASET DENGAN DAN TANPA *STEMMING*

Jumlah Klaster	**Data Tanpa *Stemming***		**Data Dengan *Stemming***	
	SSE	**Selisih**	**SSE**	**Selisih**
2	0.526	0	0.575	0
3	0.441	0.085	0.426	0.149
4	0.331	0.110	0.339	0.087
5	0.292	0.039	0.270	0.069
6	0.240	0.052	0.249	0.021
7	0.203	0.037	0.191	0.058
8	0.178	0.025	0.186	0.005
9	0.164	0.014	0.154	0.032
10	0.159	0.005	0.146	0.008
Rerata	**0.282**		**0.282**	

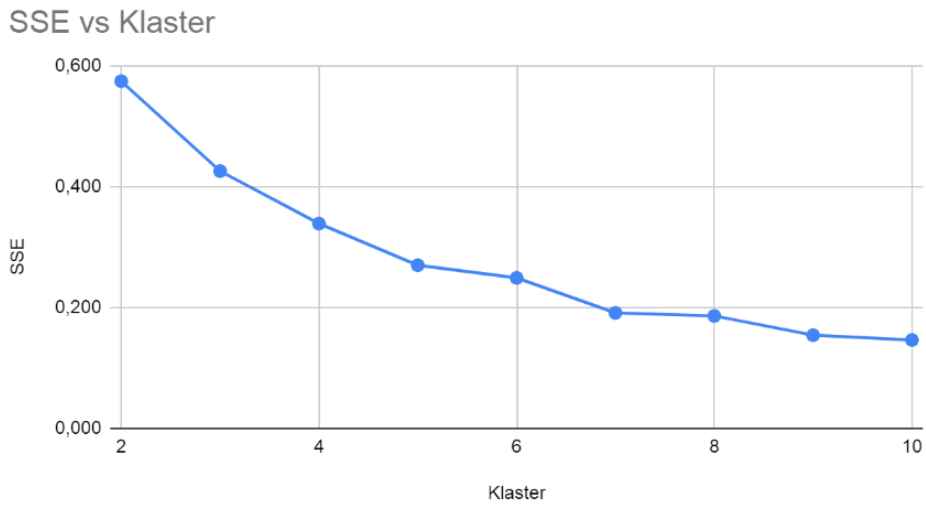
Tabel II di atas merupakan data perbandingan penggunaan metode *elbow* pada algoritma *k-means* antara dataset dengan dan tanpa menggunakan *stemming*. Untuk dataset tanpa *stemming* jumlah klaster optimalnya adalah 4 buah. Hal ini dapat dilihat dari selisih nilai SSE yang paling besar. Adapun nilai SSE pada dataset tanpa *stemming* sebesar 0.331. Di lain data, yaitu yang menggunakan *stemming*, didapatkan jumlah klaster optimalnya sebanyak 3 buah, dengan nilai SSE sebesar 0.426, di mana selisih nilai SSE sebesar 0.149. Jika dilihat nilai SSE dari kedua dataset, maka performa model optimal yang terbentuk adalah pada dataset tanpa menggunakan *stemming*. Akan tetapi, rerata SSE dari kedua dataset bernilai sama, yaitu pada angkat 0.282.

Jika dilihat lebih detail lagi dari keseluruhan percobaan, perbandingan kedua dataset dari segi nilai SSE memiliki keunggulan masing-masing. Nilai SSE terkecil pada klaster 2 diperoleh dari dataset tanpa *stemming*, yaitu sebesar 0.526 berbanding 0.575. Pada jumlah klaster 3, nilai SSE pada data dengan *stemming* lebih kecil, yaitu sebesar 0.426 berbanding 0.441. Pada jumlah klaster 4, nilai SSE terkecil terdapat pada dataset tanpa *stemming*, yaitu sebesar 0.331 berbanding 0.339, yang berarti hanya sedikit perbedaannya. Kemudian pada jumlah klaster 5, nilai SSE terkecil terdapat pada dataset dengan *stemming*, yaitu sebesar 0.270 berbanding 0.292 pada dataset tanpa *stemming*. Selanjutnya pada jumlah klaster 6, nilai SSE terkecil terdapat pada dataset tanpa *stemming* dengan nilai 0.240 berbanding 0.249. Pada jumlah klaster 7, nilai SSE terkecil diperoleh dari dataset dengan *stemming* sebesar 0.191 berbanding 0.203. Pada klaster 8 buah, dataset tanpa *stemming* lebih unggul dengan nilai SSE sebesar 0.178 berbanding 0.186. Pada klaster 9 dan 10 dataset dengan *stemming* lebih unggul dengan nilai masing-masing SSE sebesar 0.154 dan 0.146 berbanding 0.164 dan 0.159. Dapat dilihat juga bahwa semakin banyak jumlah klaster nilai SSE untuk kedua dataset memiliki selisih yang semakin sedikit. Jika digambarkan untuk masing-masing dataset maka akan menjadi seperti pada gambar 4 dan 5 berikut.



Gambar 4. Grafik metode *elbow* untuk dataset tanpa *stemming*

Gambar 4 di atas menunjukkan grafik dari hasil pengolahan menggunakan metode *elbow* pada data tanpa *stemming* di mana jumlah klaster optimalnya adalah 4 buah. Dari gambar 4 di atas juga dapat dilihat lekukan yang terlihat seperti siku juga berada pada klaster 4, hal ini menunjukkan algoritma *elbow* dapat memilih jumlah *k* yang optimal pada algoritma *k-means*. Selain itu, semakin besar jumlah klaster maka cenderung semakin kecil juga nilai SSE yang dihasilkan, hingga cenderung melandai.



Gambar 5. Grafik metode *elbow* untuk dataset dengan *stemming*

Gambar 5 di atas menunjukkan hasil pengolahan data dengan *stemming* di mana jumlah klaster optimalnya sebanyak 3 buah. Hampir sama dengan hasil olahan data tanpa *stemming*, semakin banyak jumlah klaster maka tingkat error atau SSE juga cenderung menurun pada data dengan menggunakan *stemming*.

TABEL III
PERBANDINGAN PEFORMA DBI PADA DATASET DENGAN DAN TANPA *STEMMING*

Jumlah Klaster	***Max Iteration***	**DBI dengan *Stemming***	**DBI tanpa *Stemming***
2	100	0.003	0.003
3	100	0.003	0.002

Jumlah Klaster	*Max Iteration*	DBI dengan *Stemming*	DBI tanpa *Stemming*
4	100	0.003	0.003
5	100	0.003	0.003
6	100	0.003	0.002
7	100	0.003	0.002
8	100	0.003	0.002
9	100	0.003	0.002
10	100	0.003	0.002

Pada tabel III di atas menunjukkan data perbandingan performa algoritma *k-means* dan *elbow* berdasarkan nilai DBI. Dapat dilihat bahwa baik dataset yang menggunakan maupun tanpa *stemming* menunjukkan nilai DBI yang sama, yaitu 0.003 untuk masing-masing jumlah klaster 3 dan 4. Pada dataset dengan menggunakan *stemming*, nilai DBI untuk jumlah klaster 2 sampai 10 adalah sama. Berbeda dengan dataset tanpa *stemming*, terdapat variasi nilai DBI, meskipun selisihnya sangat kecil. Pada jumlah klaster 2, 4 dan 5 memiliki nilai DBI sebesar 0.003, sedangkan selainnya memiliki nilai DBI sebesar 0.002.

Jika dilihat dari sisi konten dari masing-masing klaster yang terbentuk, dapat dijelaskan sebagai berikut. Pada dokumen *tweet* tanpa menggunakan *stemming* yang menghasilkan 4 buah klaster, klaster pertama terdiri dari 481 buah data atau 48.1% dari total data. Pada klaster pertama ini membahas tentang banyaknya siswa yang memilih untuk sekolah luring atau offline. Beberapa alasan dari keinginan tersebut adalah bahwa dengan belajar secara luring atau offline materi lebih mudah dipahami, serta pembelajaran online atau daring yang dirasa tidak efektif, sehingga banyak siswa yang mengeluh. Sebaran kata dapat dilihat pada *wordcloud* berikut ini.

Gambar 6. *Wordcloud* klaster pertama model klasterisasi dataset tanpa *stemming*

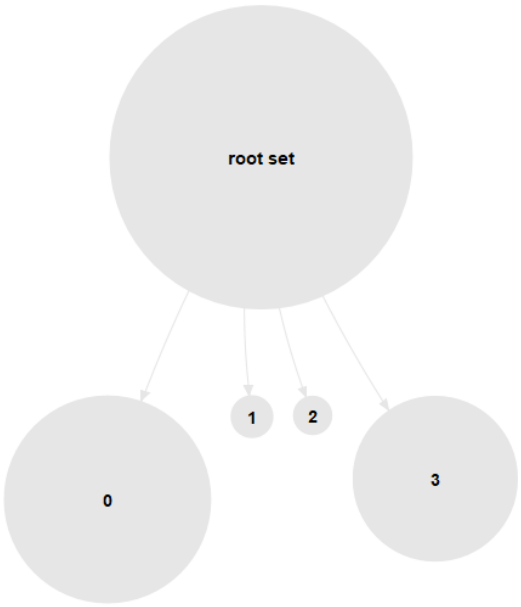
Gambar 6 di atas dapat dilihat bahwa kata kunci tatap muka dan ptm dapat terbaca dengan ukuran yang besar, dengan kata lain kata tersebut sering muncul pada klaster ini.

Klaster kedua terdiri dari 84 buah data atau sekitar 8.4% dari keseluruhan data. Pada klaster ini membahas tentang persiapan sekolah tatap muka atau sekolah offline. Salah satu yang paling

banyak dibahas adalah pencarian informasi tentang vaksinasi covid-19. Pada klaster kedua juga memiliki sebaran kata atau *wordcloud* yang relative sama dengan klaster pertama.

Klaster ketiga terdiri dari 77 buah data atau 7.7% dari total keseluruhan data. Pada klaster ketiga ini berisi tentang siswa yang sudah terlanjut nyaman dengan sekolah online. Klaster ketiga ini merupakan klaster dengan jumlah data paling sedikit untuk dataset tanpa menggunakan *stemming*.

Selanjutnya klaster keempat terdiri dari 361 buah data atau sekitar 36.1% dari total keseluruhan. Klaster ini berisi tentang pendapat masyarakat tentang pembelajaran offline yang masih berbahaya. Ada juga yang berpendapat bahwa pembelajaran offline boleh saja dilakukan dengan catatan harus dikombinasikan secara *hybrid*. Secara keseluruhan, sebaran data ke dalam klaster-klaster pada data tanpa *stemming* dapat dilihat pada gambar 7 berikut ini.



Gambar 7. Sebaran data tanpa *stemming* ke dalam masing-masing klaster

Gambar 7 di atas nampak sebaran data tanpa *stemming* bahwa klaster dengan data terbanyak terdapat pada klaster 1 dan klaster 3, di mana pada gambar tertulis klaster 0 dan 3. Sedangkan pada klaster 2 dan 3 terlihat sangat kecil karena hanya berisi 7.7% sampai 8.4% saja.

IV. KESIMPULAN

Model klasterisasi yang terbentuk dari penelitian ini pada 1000 dataset sampel dengan menggunakan *stemming* didapatkan jumlah klaster terbaik sebanyak 3 buah klaster, dengan nilai SSE sebesar 0.426 dan nilai DBI sebesar 0.003. Dari 3 buah klaster ini, pada klaster pertama membahas tentang kesiapan pembelajaran tatap muka, dan juga protokol kesehatan yang ketat, perlunya kehati-hatian ketika sudah melakukan sekolah tatap muka. Pada klaster kedua membahas tentang banyak yang memilih sekolah offline dibanding online karena lebih mudah pembelajaran tatap muka langsung. Pada klaster ketiga membahas tentang sudah terbiasanya sekolah online, sehingga perlu adaptasi untuk pembelajaran tatap muka langsung, masih cenderung memilih online karena kondisi dan sebagian yang lain masih ragu-ragu yang ditandai dengan sama-sama memilih online dan offline. Model klasterisasi yang terbentuk dari penelitian ini untuk 1000 dataset sampel tanpa menggunakan *stemming* didapatkan jumlah klaster terbaiknya sebanyak 4 buah klaster, dengn nilai SSE sebesar 0.331 dan nilai DBI sebesar

0.003. Pada klaster pertama membahas banyak siswa yang berminat untuk sekolah offline karena lebih mudah memahami dibandingkan sekolah online, salah satunya ditandai ada yang mengeluh karena terlalu lama sekolah online. Pada klaster kedua berisi tentang persiapan sekolah tatap muka yang ditandai dengan pencarian informasi tentang vaksinasi. Klaster ketiga membahas tentang siswa yang sudah terlanjut nyaman sekolah online. Pada klaster keempat membahas tentang pendapat bahwa sekolah offline masih berbahaya, namun ada juga yang menyatakan setuju sekolah tatap muka namun dipadukan dengan metode pembelajaran hybrid. Dengan demikian jika dilihat dari nilai DBI, dataset dengan dua pendekatan ini memiliki model dengan performa klasterisasi yang sama. Berbeda jika dilihat dari nilai SSE, model klasterisasi dengan dataset tanpa stemming memiliki performa lebih baik jika dibandingkan dataset yang menggunakan stemming. Kesimpulan lain yang didapatkan pada penelitian ini menunjukkan bahwa stemming kurang baik diterapkan, terutama untuk data teks yang singkat seperti yang didapatkan melalui platform media sosial *Twitter*.

REFERENSI

[1] G. Guillén, T. Sawin, and N. Avineri, "Persepsi Mahasiswa Dalam Penggunaan Teknologi Pembelajaran Bahasa Arab pada Pertemuan Tatap Muka Terbatas di Masa Pandemi COVID-19," *Alibbaa' J. Pendidik. Bhs. Arab*, vol. 3, no. 1, pp. 320–328, 2022.

[2] C. N. Rohim, R. K. Wiryaningtyas, L. N. C. L. Aji, Y. A. Priambudi, and Darmadi, "Masalah Yang Muncul Pada Pelaksanaan Kegiatan Pembelajaran Luring Di Masa Pandemi," *Innov. J. Soc. Sci. Res.*, vol. 2, no. 1, pp. 228–231, 2022.

[3] S. F. Nissa and A. Haryanto, "Implementasi Pembelajaran Tatap Muka Di Masa Pandemi Covid-19," *J. IKA PGSD (Ikatan Alumni PGSD) UNARS*, vol. 8, no. 2, pp. 402–409, 2020.

[4] A. Rangrej, S. Kulkarni, and A. V. Tendulkar, "Comparative study of clustering techniques for short text documents," in *Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011*, pp. 111–112, 2011.

[5] T. S. Kartikasari, H. Setiawan, and P. L. T. Irawan, "Implementasi Text Mining Untuk Analisis Opini Publik Terhadap Calon Presiden," *J. SimanteC*, vol. 7, no. 1, pp. 39–47, 2018.

[6] Y. Xing, X. Wang, C. Qiu, Y. Li, and W. He, "Research on opinion polarization by big data analytics capabilities in online social networks," *Technol. Soc.*, vol. 68, no. January, p. 101902, 2022.

[7] H. Irsyad, A. Farisi, and M. R. Pribadi, "Klasifikasi Opini Masyarakat Terhadap Jasa ISP MyRepublic dengan Naïve Bayes," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 1, pp. 30–34, 2019.

[8] K. Ariasa, I. G. A. Gunadi, and I. M. Candiasa, "Optimasi Algoritma Klaster Dinamis pada K-Means dalam Pengelompokkan Kinerja Akademik Mahasiswa (Studi Kasus: Universitas Pendidikan Ganesha)," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 9, no. 2, pp. 181–193, 2020.

[9] R. C. Balabantaray, C. Sarma, and M. Jha, "Document Clustering using K-Means and K-Medoids," *Int. J. Knowl. Based Comput. Syst.*, vol. 1, no. 1, pp. 7–13, 2013.

[10] Zulhanif, Sudartianto, B. Tantular, and I. G. N. M. Jaya, "Aplikasi Latent Dirichlet Allocation (Lda) Pada Clustering Data Teks," *J. Log.*, vol. 7, no. 1, pp. 46–51, 2017.

[11] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artif. Intell. Med.*, vol. 117, no. May 2020.

[12] S. Rustam, H. A. Santoso, and C. Supriyanto, "Optimasi K-Means Clustering Untuk

Identifikasi Daerah Endemik Penyakit Menular Dengan Algoritma Particle Swarm Optimization Di Kota Semarang," *Ilk. J. Ilm.*, vol. 10, no. 3, pp. 251–259, 2018.

[13] D. P. Isnarwaty and Irhamah, "Text clustering pada akun twitter layanan ekspedisi JNE , J&T, dan Pos Indonesia menggunakan metode Density-Based Spatial Clustering of Applications with Noise (DBSCAN)," *J. Sains dan Seni*, vol. 8, no. 2, pp. 137–144, 2019.

[14] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 918, 2019.

[15] D. Merlini and M. Rossini, "Text categorization with WEKA: A survey," *Mach. Learn. with Appl.*, vol. 4, no. April, p. 100033, 2021.

[16] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, p. 100008, 2021.

[17] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognit.*, vol. 93, pp. 95–112, 2019.

[18] A. F. Febrianti, A. H. Cabral, and G. Anuraga, "K-Means Clustering Dengan Metode Elbow Untuk Pengelompokan Kabupaten Dan Kota Di Jawa Timur," *Semin. Nas. Has. Ris. dan Pengabdi. -SNHRP*, pp. 863–870, 2018.

[19] A. F. Hadi, D. Bagus, and M. Hasan, "Text Mining Pada Media Sosial Twitter Studi Kasus : Masa Tenang Pilkada DKI 2017 Putaran 2," in *Seminar Nasional Matematika dan Aplikasinya, 21 Oktober 2017 Surabaya, Universitas Airlangga*, pp. 324–331, 2017.

[20] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Expert Syst. Appl.*, vol. 172, no. May 2020, p. 114652, 2021.

[21] E. Supriyadi, A. Basuki, and R. Sigit, "Klasterisasi Kualitas Beras Berdasarkan Citra Pecahan Bulir Dan Sebaran Warna," *J. INOVTEK POLBENG - SERI Inform.*, vol. 6, no. 1, pp. 105–119, 2021.

[22] S. F. Susilo, A. Jamaludin, and I. Purnamasari, "Pengelompokan Desa Menggunakan K-Means Untuk Penyelenggaraan Penanggulangan Bencana Banjir," *JOINS (Journal Inf. Syst.*, vol. 5, no. 2, pp. 156–167, 2020.

[23] K. Thirumoorthy and K. Muneeswaran, "A hybrid approach for text document clustering using Jaya optimization algorithm," *Expert Syst. Appl.*, vol. 178, no. April, pp. 1–16, 2021.

[24] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 1, pp. 48–53, 2020.

[25] D. Zhao *et al.*, "k-means clustering and kNN classification based on negative databases," *Appl. Soft Comput.*, vol. 110, p. 107732, 2021.

UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi Republik Indonesia yang telah memberikan pendanaan penelitian ini melalui DRTPM Kemdikbudristek Tahun 2022. Penulis juga mengucapkan terima kasih kepada Tim *Jurnal Informatika Polbeng* yang telah meluangkan waktu untuk membuat template ini, termasuk pemeriksaan sampai diterbitkannya artikel ini.