PENGELOMPOKKAN BERITA KESEHATAN PADA SOSIAL MEDIA TWITTER DENGAN METODE K-MEANS CLUSTERING

ANITA[1], JEFFER OLIANDO[2]
Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
jefferoliando@yahoo.com[1], anitayakub_pilchan@yahoo.com[2]

Abstract: *Twitter is one of the most influential social media in the world with 15.7 million users in Indonesia and is ranked 6th as the country with the most Twitter users. Besides being used as social media, Twitter is also used as a medium to read or send news. News information that continues to increase causes users to have difficulty in finding certain news information. One solution that can be applied to overcome this problem is through clustering of news information on Twitter. In this study, the researcher used quantitative research with the K-Means Clustering method, which is one of the clustering methods used in grouping data. The data used in this study is a dataset taken from the UCI Machine Learning Repository, namely Health News in Twitter Data Set as many as 14,970 tweet data. The results showed that the determination of the best cluster using the Elbow method on the dataset resulted in empirical evidence that the best cluster was K=3. The results of grouping health news on Twitter social media using the K-Means Clustering method with K=3 resulted in the number of clusters, namely C1 as many as 4,991 data tweets, C2 as many as 4,482 tweets data, and C3 as many as 5,497 tweets.*
Keywords: *Health News Grouping; Twitter Social Media; Elbow Method; K-Means Clustering Method.*

Abstrak: Twitter merupakan salah satu media sosial paling berpengaruh di dunia dengan 15,7 juta pengguna di Indonesia dan menduduki peringkat ke-6 sebagai negara dengan pengguna Twitter terbanyak. Selain digunakan sebagai media sosial, Twitter juga digunakan sebagai media untuk membaca atau mengirim berita. Informasi berita yang terus meningkat menyebabkan pengguna kesulitan dalam mencari informasi berita tertentu. Salah satu solusi yang dapat diterapkan untuk mengatasi permasalahan tersebut adalah melalui clustering informasi berita di Twitter. Dalam penelitian ini peneliti menggunakan penelitian kuantitatif dengan metode K-Means Clustering yang merupakan salah satu metode clustering yang digunakan dalam pengelompokan data. Data yang digunakan dalam penelitian ini adalah dataset yang diambil dari UCI Machine Learning Repository yaitu Health News in Twitter Data Set sebanyak 14.970 data tweet. Hasil penelitian menunjukkan bahwa penentuan cluster terbaik menggunakan metode Elbow pada dataset menghasilkan bukti empiris bahwa cluster terbaik adalah K=3. Hasil pengelompokan berita kesehatan di media sosial Twitter menggunakan metode K-Means Clustering dengan K=3 menghasilkan jumlah cluster yaitu C1 data tweet sebanyak 4.991, C2 data tweet sebanyak 4.482, dan C3 sebanyak 5.497 tweet.
Kata Kunci: Pengelompokan Berita Kesehatan; Media Sosial Twitter; Metode Siku; Metode Pengelompokan K-Means.

A.Pendahuluan

Seluruh Berita merupakan salah satu sumber informasi terkini yang ada pada media massa seperti surat kabar, televisi, dan media *online*. Seiring berkembangnya internet yang memberikan kemudahan untuk mendapatkan suatu informasi berita salah satunya melalui sosial media (Kasanah, Muladi, & Pujianto, 2019). Salah satu sosial media yang paling berpengaruh di dunia adalah Twitter. Di Indonesia terdapat sebanyak 15,7 juta pengguna Twitter dan Indonesia menjadi negara peringkat ke 6 sebagai negara dengan pengguna Twitter terbanyak di bawah Amerika Serikat, Brazil, dan negara-negara lainnya (Dihni & Bayu, 2021). Twitter menjadi salah satu sosial media yang paling digemari dan sering dijadikan sebagai tempat untuk berbagi informasi, baik itu informasi berupa komentar ataupun pesan (Nugroho, 2018). Selain digunakan sebagai media sosial, Twitter juga digunakan sebagai media untuk membaca ataupun mengirim suatu berita. Informasi berita yang terus meningkat menyebabkan

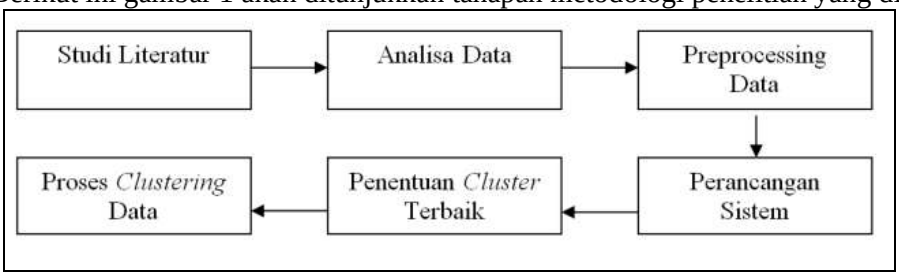
 P-ISSN 2622-9110
E-ISSN 2654-8399

pengguna akan mengalami kesulitan dalam mencari informasi-informasi berita tertentu (Kurniawan, Fauzi, & Widodo, 2017).

Salah satu solusi yang dapat diterapkan untuk dapat mengatasi masalah tersebut adalah melalui *clustering* terhadap informasi berita yang ada di Twitter. *Clustering* merupakan proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat (Madhulatha, 2012). Banyaknya berita yang tersedia di Twitter perlu dikelompokkan agar dapat memudahkan pengguna dalam melakukan pencarian berita. Metode ini dipilih pada penelitian ini karena tergolong cukup akurat serta memiliki waktu proses komputasi yang relatif singkat (Pramudita, Putro, & Makhmud, 2018).

B.Metodologi Penelitian

Pada penelitian ini, peneliti menggunakan jenis penelitian secara kuantitatif dengan metode *K-Means Clustering*. Penelitian kuantitatif adalah penelitian ilmiah yang sistematis terhadap bagian-bagian dan fenomena serta kualitas hubungan-hubungannya. Tujuan penelitian kuantitatif adalah mengembangkan dan menggunakan model-model matematis, teori-teori atau hipotesis yang berkaitan dengan fenomena alam (Sugiyono, 2017). Metode *K-Means Clustering* merupakan metode klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster *(centroid)* terdekat dengan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antara klaster. Penelitian ini menerapkan metode *K-Means Clustering* dalam mengelompokkan berita kesehatan Twitter agar dapat memudahkan pengguna Twitter dalam mencari informasi-informasi berita tertentu. Selain itu juga akan ditentukan klaster terbaik dari pengujian yang dilakukan sehingga hasil pengujian dapat dijadikan acuan dalam mennetukan klaster berita yang paling tepat bagi pengguna Twitter. Berikut ini gambar 1 akan ditunjukkan tahapan metodologi penelitian yang dilakukan.



Gambar 1. Tahapan Metolodogi Penelitian

C.Hasil dan Pembahasan

Peralatan yang digunakan pada penelitian ini berupa laptop dengan prosesor Intel® Core™ i3-4005U CPU @1,70GHz dan RAM 2 GB, serta menggunakan sistem operasi Windows 10. Sistem pengelompokkan berita kesehatan pada penelitian ini dibangun dengan menggunakan bahasa pemrograman Python versi 3.9.6 dengan text editor yang digunakan adalah Python IDLE (Integrated DeveLopment Environment). Output hasil clustering akan ditampilkan menggunakan Python Shell.

Hasil Analisa Data

Semua Pada tahapan ini akan dilakukan analisis eksperimental dengan mengamati dataset yang digunakan pada penelitian. Dataset pada penelitian ini berformat .txt dan dipisahkan dengan huruf '|' di setiap atributnya. Setiap *record* dipisah melalui baris baru sehingga dapat disimpulkan satu baris adalah satu *record*. Setelah peneliti mengamati dataset, secara garis besar peneliti memilih 4 jenis dataset dengan kriteria sebagai berikut:

1. Dataset 1 merupakan *tweet* dari BBC Health News dengan jumlah data di dalamnya sebanyak 3.929 *record* data.
2. Dataset 2 merupakan *tweet* dari CBC Health News dengan jumlah data di dalamnya sebanyak 3.741 *record* data.

3.Dataset 3 meruapakn *tweet* dari CNN Health News dengan jumlah data di dalamnya sebanyak 4.061 *record* data.
4.Dataset 4 merupakan *tweet* dari Everyday Health News dengan jumlah data di dalamnya sebanyak 3.239 *record* data.

Berikut ini pada gambar 2 akan ditampilkan contoh dataset yang digunakan pada penelitian ini.

Gambar 2. Contoh Dataset Penelitian

Hasil Preprocessing Data

Keempat dataset pada penelitian ini belum dapat langsung dipakai karena masih perlu dilakukan pembersihan *(cleaning)* data. Berikut ini tabel 1 menunjukkan hasil dari preprocessing data yang dilakukan.

Tabel 1. Hasil Preprocessing Data

Tahapan	Data Sebelum Preprocessing	Data Sesudah Preprocessing
Menghapus \n di akhir setiap kalimat.	574965711967223808|Mon Mar 09 16:11:22 +0000 2015|RT @HaertlG: Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from @WHO: http://t.co/CDfeyjE3k…	574965711967223808|Mon Mar 09 16:11:22 +0000 2015|RT @HaertlG: Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from @WHO: http://t.co/CDfeyjE3k…
Menghapus ID Tweet dan stempel waktu.	574965711967223808|Mon Mar 09 16:11:22 +0000 2015|RT @HaertlG: Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from @WHO: http://t.co/CDfeyjE3k…	RT @HaertlG: Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from @WHO: http://t.co/CDfeyjE3k…
Menghapus semua kata yang dimulai dengan simbol @.	RT @HaertlG: Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany.	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in

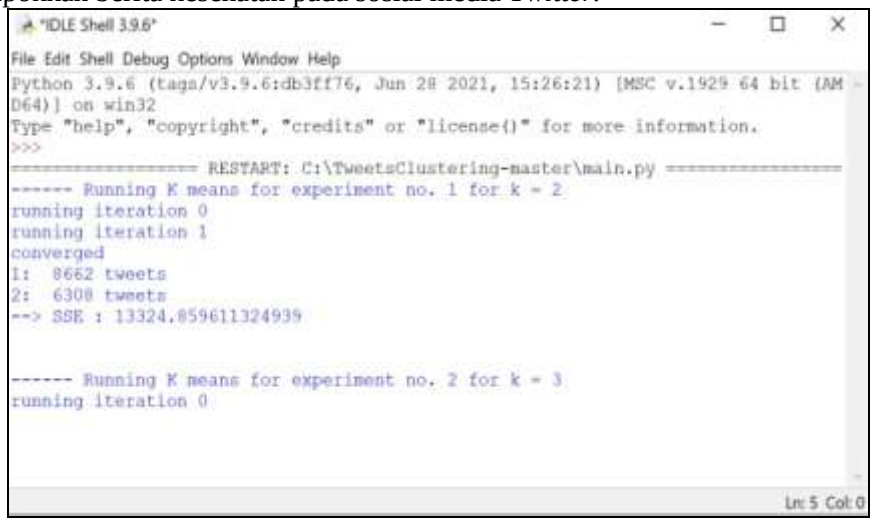
P-ISSN 2622-9110
E-ISSN 2654-8399

	Here is the posting from @WHO: http://t.co/CDfeyjE3k…	Germany. Here is the posting from http://t.co/CDfeyjE3k…
Menghapus URL	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from http://t.co/CDfeyjE3k…	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from
Menghapus titik dua dari akhir kata	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from
Menghapus semua simbol tagar *(hash tags)*.	RT Middle East respiratory syndrome coronavirus (#MERS-CoV) case in Germany. Here is the posting from	RT Middle East respiratory syndrome coronavirus (MERS-CoV) case in Germany. Here is the posting from
Mengubah setiap kata menjadi huruf kecil *(case folding)*.	RT Middle East respiratory syndrome coronavirus (MERS-CoV) case in Germany. Here is the posting from	rt middle east respiratory syndrome coronavirus (mers-cov) case in germany. here is the posting from
Menghapus tanda baca	rt middle east respiratory syndrome coronavirus (mers-cov) case in germany. here is the posting from	rt middle east respiratory syndrome coronavirus merscov case in germany here is the posting from
Menghapus spasi berlebih	rt middle east respiratory syndrome coronavirus merscov case in germany here is the posting from	rt middle east respiratory syndrome coronavirus merscov case in germany here is the posting from

Tabel 1 menunjukkan ilustrasi tahapan preprocessing data teks pada salah *tweets* judul berita yang akan dikaji. Hasil preprocessing data teks ini mampu menghapus jumlah kata tidak bermakna dari keseluruhan kata yang ada pada data awal. Hal ini tentunya sangat membantu dalam melakukan kajian data teks pada analisis lanjutan.

Hasil Perancangan Sistem

Berikut ini gambar 3 akan ditampilkan hasil dari rancangan sistem untuk mengelompokkan berita kesehatan pada sosial media *Twitter.*



Gambar 3. Hasil Rancangan Sistem

Hasil Penentuan *Cluster* Terbaik

Pada tahapan ini penentuan *cluster* terbaik akan ditentukan dengan menggunakan metode *Elbow*. Untuk mencari *cluster* terbaik (nilai k) yang optimal maka nilai k akan dicek satu persatu dan akan dicatat nilai *SSE (Sum Square Error)*. Berikut ini tabel 2 menunjukkan hasil SSE dari tiap-tiap cluster dari eksperimen pertama dengan data sebanyak 3.000 *record tweets*.

Tabel 2. Hasil SSE Dari Setiap *Cluster* Untuk 3.000 *Record Tweets*

Klaster	SSE	Selisih
2	2.609,18	0
3	2.574,15	35,03
4	2.563,86	10,29
5	2.569,17	-5,31
6	2.504,35	64,82
7	2.447,76	56,59
8	2.473,58	-25,82

Gambar 4. Grafik SSE 3.000 *Record Tweets*

Dari hasil proses perhitungan SSE terhadap 3.000 *record tweets* maka hasil yang mengalami penurunan yang paling besar adalah pada K=6. Selanjutnya akan dilakukan eksperimen kedua dengan 6.000 *record tweets.* Berikut ini tabel 3 menunjukkan hasil SSE dari tiap-tiap cluster dari eksperimen kedua dengan data sebanyak 6.000 *record tweets*.

Tabel 3. Hasil SSE Dari Setiap *Cluster* Untuk 6.000 *Record Tweets*

Klaster	SSE	Selisih
2	5.290,94	0
3	5.182,64	108,30
4	5.228,86	-46,22
5	5.083,07	145,79
6	5.052,65	30,42
7	5.040,29	12,36
8	5.017,63	22,66

Gambar 5. Grafik SSE 6.000 *Record Tweets*

P-ISSN 2622-9110
E-ISSN 2654-8399

Berbeda dengan eksperimen pertama, hasil proses perhitungan SSE dari eksperimen kedua terhadap 6.000 *record tweets* mendapatkan hasil yang mengalami penurunan yang paling besar adalah pada K=5. Selanjutnya akan dilakukan eksperimen ketiga dengan 9.000 *record tweets.* Berikut ini tabel 4 menunjukkan hasil SSE dari tiap-tiap cluster dari eksperimen ketiga dengan data sebanyak 9.000 *record tweets*.

Tabel 4. Hasil SSE Dari Setiap *Cluster* Untuk 9.000 *Record Tweets*

Klaster	SSE	Selisih
2	8.050,98	0
3	7.856,96	194,02
4	7.920,92	-63,96
5	7.738,65	182,27
6	7.697,07	41,58
7	7.741,33	-44,26
8	7.671,59	69,74

Gambar 6. Grafik SSE 9.000 *Record Tweets*

Berbeda dengan eksperimen pertama dan kedua, hasil proses perhitungan SSE dari eksperimen ketiga terhadap 9.000 *record tweets* mendapatkan hasil yang mengalami penurunan yang paling besar adalah pada K=3.

Selanjutnya akan dilakukan eksperimen keempat dengan 12.000 *record tweets.* Berikut ini tabel 3.5 menunjukkan hasil SSE dari tiap-tiap cluster dari eksperimen keempat dengan data sebanyak 12.000 *record tweets*.

Tabel 5. Hasil SSE Dari Setiap *Cluster* Untuk 12.000 *Record Tweets*

Klaster	SSE	Selisih
2	10.899,34	0
3	10.540,83	358,51
4	10.385,23	155,60
5	10.410,54	-25,31
6	10.229,04	181,50
7	10.246,34	-17,30
8	10.141,43	104,91

Gambar 7. Grafik SSE 12.000 *Record Tweets*

Hasil proses perhitungan SSE dari eksperimen keempat terhadap 12.000 *record tweets* mendapatkan hasil yang sama dengan eksperimen ketiga dimana penurunan yang paling besar adalah pada K=3. Selanjutnya akan dilakukan eksperimen kelima dengan 14.970 *record tweets.* Berikut ini tabel 6 menunjukkan hasil SSE dari tiap-tiap cluster dari eksperimen kelima dengan data sebanyak 14.970 *record tweets*.

Tabel 6. Hasil SSE Dari Setiap *Cluster* Untuk 14.970 *Record Tweets*

Klaster	SSE	Selisih
2	13.321,07	0
3	13.014,84	306,23
4	12.840,87	173,97
5	12.778,69	62,18
6	12.646,38	132,31
7	12.562,64	83,74
8	12.591,14	-28,50

Gambar 8. Grafik SSE 14.970 *Record Tweets*

Hasil proses perhitungan SSE dari eksperimen kelima terhadap 14.970 *record tweets* mendapatkan hasil yang sama dengan eksperimen ketiga dan keempat dimana penurunan yang paling besar adalah pada K=3. Jadi dari hasil eksperimen dapat disimpulkan bahwa jumlah klaster yang ideal adalah K=3 dikarenakan terdapat 3 eksperimen yang mengalami penurunan nilai SSE paling besar pada klaster tersebut. K = 3 akan dijadikan *default* klaster untuk menentukan karakteristik dari data-data *tweets* berita kesehatan pada penelitian ini.

Hasil *Clustering* Data

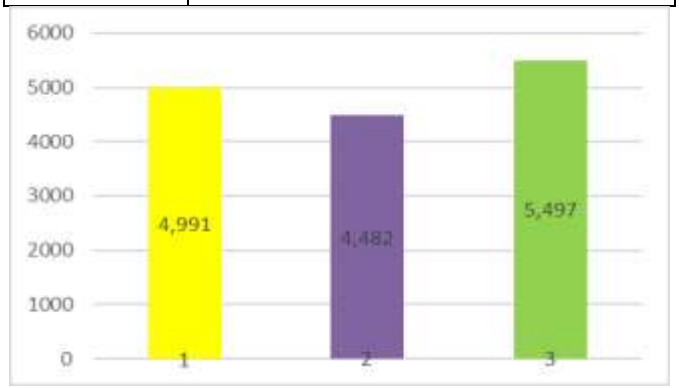
Pada tahapan ini akan ditunjukkan hasil dari pengelompokkan berita kesehatan pada sosial media *Twitter* menggunakan metode *K-Means Clustering*. Berdasarkan tahapan sebelumnya didapatkan nilai K terbaik dan yang paling ideal adalah 3 sehingga nilai K yang

P-ISSN 2622-9110
E-ISSN 2654-8399

digunakan dalam proses *clustering* data adalah K=3. Berikut ini tabel 7 menunjukkan hasil *clustering* data.

Tabel 7**.** Hasil *Clustering* Data

Klaster	Jumlah Anggota Klaster (*Tweets*)
1	**4.991**
2	**4.482**
3	**5.497**



Gambar 9. Grafik Hasil *Clustering* Data

Pembahasan

Dari kelima hasil percobaan pada dataset dengan metode *Elbow*, penelitian ini menghasilkan bukti empiris bahwa klaster terbaik adalah K=3. Berdasarkan hasil eksperimen yang dilakukan didapatkan 3 dari 5 eksperimen mendapatkan penurunan nilai SSE yang paling besar pada K=3. Hasil penelitian ini sesuai dengan penelitian yang dilakukan oleh (Larasati, Maren, & Wulandari, 2021) dimana penelitian yang dilakukan mencari klaster terbaik dari isi konten *tweets* sosial media Indonesia. Hasil penelitian menunjukkan bahwa klaster terbaik adalah K=3. Dari tabel 7 dan grafik 9 terlihat bahwa proses pengelompokkan berita kesehatan pada sosial media *Twitter* pada K=3 menghasilkan jumlah anggota klaster terbesar adalah klaster 3 sebanyak 5.497 *tweets*, kedua terbesar adalah klaster 1 sebanyak 4.991 *tweets*, dan jumlah anggota klaster yang paling kecil adalah klaster 2 sebanyak 4.482 *tweets*.

D.Penutup

Berikut ini akan diuraikan kesimpulan dari hasil penelitian yang telah dilakukan yaitu: 1) Penentuan jumlah *cluster* terbaik dengan menggunakan metode *Elbow* belum tentu menghasilkan jumlah *cluster* K yang sama pada jumlah data yang berbeda-beda; 2) Implementasi metode *Elbow* pada penelitian ini dengan jumlah eksperimen sebanyak 5 kali pada jumlah dataset yang berbeda menghasilkan jumlah *cluster* terbaik adalah K=3;dan 3) Hasil pengelompokkan berita kesehatan pada sosial media *Twitter* dengan K=3 menghasilkan jumlah *cluster* yaitu C1 sebanyak 4.991 data *tweets*, C2 sebanyak 4.482 data *tweets*, dan C3 sebanyak 5.497 *tweets*.

Daftar Pustaka

Dihni, V. A., & Bayu, D. J. (2021). *Inilah 10 Negara dengan Pengguna Twitter Terbanyak, Ada Indonesia?* Retrieved November 4, 2021, from https://databoks.katadata.co.id/datapublish/2021/11/04/inilah-10-negara-dengan-pengguna-twitter-terbanyak-ada-indonesia

Kasanah, A. N., Muladi, & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Classdalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI), III*(2), 196-201.

Kurniawan, B., Fauzi, M. A., & Widodo, A. W. (2017). Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, I*(10), 1193-1200.

Kom, HARIYADI S. "Pengembangan Sistem Pakar Berbasis Aturan Untuk Menentukan Mata Kuliah Yang Akan Diambil Ulang (Remedial) Dengan Metode Forward Chaining." *Menara Ilmu* 10.60-65 (2016).

Larasati, A., Maren, R., & Wulandari, R. (2021). Utilizing Elbow Method for Text Clustering Optimization in Analyzing Social Media Marketing Content of Indonesian e-Commerce. *Jurnal Teknik Industri, XXIII*(2), 111-119.

Madhulatha, T. S. (2012). An Overview on Clustering Methods. *IOSR Journal of Engineering*, *II*(4), 719-725.

Nugroho, A. (2018). Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstrasi Fitur N-Gram. *Jurnal Sains Komputer & Informatika (J-SAKTI), II*(2), 200-209.

Pramudita, Y. D., Putro, S. S., & Makhmud, N. (2018). Klasifikasi Berita Olahraga Menggunakan Metode Naive Bayes Dengan Enchanted Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), V*(3), 269-276.

Purba, N., Poningsih, & Tambunan, H. S. (2021). Penerapan Algoritma K-Means Clustering Pada Penyebaran Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Provinsi Riau. *Journal of Information System Research (JOSH), II*(3), 220-226.

Rachman, D. A., Goejantoro, R., & Amijaya, F. D. (2020). Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering. *Jurnal EKSPONENSIAL, XI*(2), 167-174.

Sugiyamto, Surarso, B., & Sugiharto, A. (2021). Analisa Performa Metode Cosine dan Jacard Pada Pengujian Kesamaan Dokumen. *Jurnal Masyarakat Informatika, V*(10), 1-8.

Sugiyono. (2017). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D.* Bandung: CV. Alfabeta.

 P-ISSN 2622-9110
E-ISSN 2654-8399