Clustering Artikel pada Portal Berita Online Menggunakan Metode K-Means

Dwiki Krisnanda Wardy[a1], I Ketut Gede Darma Putra[a2], Ni Kadek Dwi Rusjayanthi[a3]
[a]Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Bukit Jimbaran, Bali, Indonesia, telp. (0361) 701806
e-mail: [1]krisnanda.wardy@student.unud.ac.id ,[2]ikgdarmaputra@unud.ac.id,
[3]dwi.rusjayanthi@unud.ac.id

Abstrak

Kategori berita pada portal berita yang begitu beragam mengakibatkan kinerja para editor semakin banyak. Banyaknya artikel berita setiap bulannya, menambah tugas editor untuk mengkategorikan artikel secara manual ke dalam kategori yang telah ditentukan. Clustering dapat digunakan untuk melakukan proses pengelompokan data sehingga nantinya dapat mengelompokkan data dalam kategori yang sama dengan data yang sejenis. K-Means merupakan salah satu metode yang dapat digunakan untuk melakukan Clustering. K-Means merupakan teknik Clustering berbasis jarak yang dibagi menjadi serangkaian *cluster* dan hanya berfungsi untuk atribut numerik. Pengujian K-Means yang dilakukan dalam penelitian ini ditujukan untuk membandingkan nilai *cluster.* K-Means yang dibuat dalam penelitian ini menerapkan TF-IDF, *feature selection* dan PCA. Proses penilaian nilai *cluster* menggunakan visualisasi berupa *bar plot* dari tiap nilai *metric* yang diperhatikan yaitu *mean silhouette, accuracy, precision, recall, F1-score* dan *silouette score*. Hasil penelitian yang telah dilakukan Metode K-Means mampu mencapai 94.93% accuracy dan recall, 95.07% precision serta 94.94% F1-score.

Kata kunci: *Artikel, Clustering, K-Means, Portal Berita Online*

Abstract

The news categories on news portals are so diverse that the performance of the editors is increasing. The number of news articles each month, adds to the editor's task to manually categorize articles into predetermined categories. Clustering can be used to group data so that later it can group data in the same category with similar data. K-Means is a method that can be used to perform clustering. K-Means is a distance-based clustering technique that is divided into a series of clusters and only works for numeric attributes. The K-Means test conducted in this study is intended to compare cluster values. The K-Means made in this study apply TF-IDF, feature selection, and PCA. The cluster value assessment process uses visualization in the form of a bar plot of each metric value that is considered, namely the mean silhouette, accuracy, precision, recall, F1-score, and silhouette score. The results of the research that has been carried out by the K-Means method can achieve 94.93% accuracy and recall, 95.07% precision, and 94.94% F1-score.

Keywords: *Articles, Clustering, K-Means, Online News Portals*

1. Pendahuluan

Perubahan teknologi informasi saat ini meningkat dengan pesat. Perubahan ini menyebabkan kecenderungan masyarakat untuk mengakses informasi khususnya berita melalui media digital meningkat. Peningkatan akses berita melalui media digital mengakibatkan jumlah dokumen menjadi banyak, sehingga pencarian didalam dokumen berbasis teks menjadi tidak efektif. Permasalahan terkait pencarian didalam dokumen berbasis teks direspon dengan penelitian dibidang pemrosesan dokumen teks berbahasa indonesia. Berita yang ditampilkan pada portal berita terdiri dari beberapa kategori, meliputi berita tentang teknologi, kesehatan, olahraga dan lainnya (sebagai contoh pada portal berita kompas.com, liputan6.com, dan lain

sebagainya). Berita yang terbit setiap bulannya mencapai 400 artikel berita dari berbagai kategori artikel pada portal berita. Hal ini mengakibatkan kinerja editor semakin meningkat, dimana mengharuskan mengharuskan editor untuk mengedit artikel di saluran yang berbeda dan secara manual mengkategorikan artikel ke dalam beberapa kategori yang telah ditentukan [1].

Proses Clustering dapat membantu pengelompokan kategori berita yang dikerjakan secara manual menjadi lebih mudah dan lebih cepat. K-Means merupakan salah satu algoritma Clustering yang umum digunakan dalam proses Clustering. Kelebihan dari K-Means yaitu sederhana, efisien dan mudah diimplementasikan. Kajian terkait penggunaan K-Means Clustering untuk pengelompokkan ayat Al-Quran terjemahan Bahasa Indonesia. Penelitian ini dilakukan dengan menggunakan tahap *preprocessing* data teks, pembobotan kata dengan TF-IDF, pengelompokan data dengan K-Means serta pelabelan data untuk kata kunci. Sistem yang dihasilkan dapat menampilkan ayat dalam kelompok yang sesuai dengan kata kunci. Hasil pengujian dengan menggunakan indeks siluet Surat Al Fatihah memberikan nilai positif sebesar 0,336. Artinya data tersebut berada pada kelompok yang tepat, sedangkan dari frekuensi kata kunci dibandingkan dengan jumlah data menghasilkan presentase 53% yang berarti bahwa kata kunci merepresentasikan setengah dari data dalam *cluster* [2].

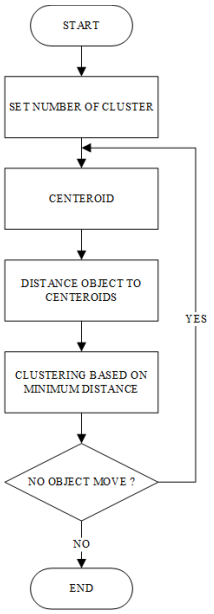
Penelitian lainnya yang terkait yaitu penerapan clustering yang digunakan pada sebuah aplikasi pendeteksi kemiripan dokumen teks Bahasa Indonesia. Teknologi informasi yang mengukur tingkat plagiarisme dokumen teks berhubungan dengan pencarian informasi dari sejumlah besar data. Memerlukan waktu yang lama untuk memproses hasil kemiripan dari seluruh isi dokumen teks. Untuk mengatasi permasalahan tersebut dibuat aplikasi yang mampu mengenali kemiripan dokumen teks dengan menggunakan teknik Clustering. Berdasarkan hasil pengujian aplikasi yang telah dijalankan, aplikasi Clustering dapat mengurangi waktu pengukuran kemiripan, namun penerapan teknik ini kurang akurat dibandingkan tanpa menerapkan teknik Clustering [3].

2. Metode Penelitian

Metode penelitian merupakan tahapan yang menjelaskan mengenai gambaran umum, dari Metode K-Means yang dipergunakan dan fase dalam proses Clustering artikel berita yang digunakan dalam penelitian.

2.1. Gambaran Umum K-Means

Metode Clustering yang digunakan dalam penelitian ini adalah K-Means. Gambaran dari proses K-Means yang digunakan dalam penelitian ini dijabarkan dalam Gambar 1.



Gambar 1. Gambaran Umum K-Means

Proses K-Means dimulai dengan menentukan jumlah *cluster*. Jumlah *cluster* digunakan untuk menentukan jumlah grup yang akan dicari. Jumlah *cluster* yang digunakan pada penelitian ini sama dengan jumlah kategori berita yang digunakan. Proses selanjutnya adalah menentukan *centroid*, dimana proses penentuan *centroid* ditentukan secara *random* dari *data point.* Proses setelah menentukan *centroid* adalah menentukan jarak dari setiap *data point* ke *centroid*. Proses selanjutnya adalah menentukan kelompok data berdasarkan jarak minimum dari setiap *data point* ke *centroid*. Proses sebelumnya kecuali proses penentuan jumlah *cluster* terus berulang sampai tidak terjadi perubahan pada *centroid*.

3. Studi Literatur

Studi literatur dilakukan dengan pengumpulan data dan informasi untuk penelitian dari berbagai sumber. Penjelasan terkait dengan Text Mining, berita online, K-Means, TF-IDF, ANOVA, PCA dan *evaluation metric* adalah sebagai berikut.

3.1. Text Mining

Text Mining juga dikenal sebagai penambangan data teks atau pengambilan pengetahuan dalam basis data teks. Jenis pekerjaan Text Mining termasuk klasifikasi, pengelompokan teks, ekstraksi konsep/entitas, analisis sentimen, peringkasan dokumen, dan entity-relation modeling (yaitu, mempelajari hubungan antar entitas). Kumpulan teks yang tidak terstruktur atau setidaknya semi terstruktur merupakan sumber data yang digunakan dalam Text Mining. Mendapatkan informasi yang berguna dari sekumpulan dokumen merupakan tujuan utama dari Text Mining [4].

3.2. Berita Online

Berita online merupakan bagian dari bentuk media online. Berita online sendiri adalah sebuah *website* yang menyediakan informasi terkini (harian) tentang satu atau lebih peristiwa yang mempengaruhi kehidupan kita sehari-hari meliputi bidang pendidikan, olahraga dan politik.

Berita online memiliki beberapa keunggulan dan kekurangan dibandingkan media cetak maupun elektronik. Keunggulan berita online adalah berita online dapat memuat informasi dalam format teks, audio, video maupun foto secara bersamaan, serta informasi yang diberikan dapat disajikan dengan mudah dan cepat, sehingga berita online juga memuat informasi yang tepat waktu. Kelemahan berita online antara lain mengabaikan keakuratan berita. Hal ini dikarenakan berita online mengutamakan kecepatan, dan berita yang dipublikasikan tidak seakurat berita media cetak, terutama dalam hal menulis dan mengandalkan perangkat computer, gadget serta koneksi internet dalam proses pembuatannya [5].

3.3. K-Means

K-Means adalah teknik Clustering berbasis jarak yang membagi data menjadi serangkaian *cluster* dan bekerja hanya dengan atribut numerik. Algoritma K-Means menyertakan partisi pengelompokan yang membagi data menjadi k subkawasan terpisah. K-Means terkenal karena kesederhanaannya dan kemampuannya untuk mengelompokkan data yang besar serta *outlier* dengan sangat cepat. Semua data harus menjadi bagian dari *cluster* tertentu, dan semua data milik *cluster* tertentu dalam satu fase proses dapat dipindahkan ke *cluster* lain oleh algoritma K-Means di fase berikutnya.

Penggunaan algoritma K-Means bersifat acak dan oleh karena itu sangat sensitif untuk menginisialisasi pusat *cluster*. Algoritma K-Means menggunakan nilai mean sebagai pusat *cluster*. Prosedur untuk algoritma K-Means dimulai dengan memilih secara acak nilai k sebagai pusat cluster awal. Proses selanjutnya membagi data yang ada menjadi k cluster untuk selanjutnya menggunakan *euclidean distance* untuk mendapatkan pusat *cluster* dan hitung kembali pusat setiap *cluster* berdasarkan rata-rata *cluster* yang diperoleh. Proses K-Means terus berulang selama ada perubahan pada grup *cluster* [6].

3.4. Term Frequency-Inverse Document Frequency (TF-IDF)

Data yang telah melewati tahap preprocessing harus numerik. Metode pembobotan TF-IDF merupakan salah satu metode yang dapat digunakan untuk mengubah data menjadi numerik. Term Frequency Invers Document Frequency (TF-IDF) adalah metode menghitung bobot kata *(term)* dalam dokumen. TF-IDF juga dikenal sebagai cara yang efisien dan memiliki hasil yang sederhana serta akurat. Metode TF-IDF ini adalah kombinasi dari dua konsep, yaitu frekuensi terbalik dari suatu dokumen yang berisi kata-kata dalam dokumen dan kata-katanya. Frekuensi kemunculan suatu kata dalam dokumen tertentu menunjukkan betapa pentingnya kata tersebut dalam dokumen tersebut [7].

3.5. ANOVA

Analysis of Variance (ANOVA) merupakan teknik analisis multivariat yang menggunakan ANOVA untuk membedakan *mean* dari dua *dataset* dengan membandingkan variansinya [8]. ANOVA merupakan salah satu metode yang dapat digunakan untuk melakukan seleksi fitur. ANOVA adalah metode statistik yang digunakan untuk menguji pengaruh signifikan antara variabel kelompok rata-rata.

3.6. PCA

Principal Component Analysis (PCA) adalah metode yang dapat digunakan untuk pengurangan dimensi. PCA adalah salah satu cara paling umum untuk mengurangi dimensi kumpulan data. PCA merupakan teknik analisis multivariat berbasis transformasi linier yang banyak digunakan untuk mereduksi dimensi data. PCA juga digunakan untuk mengekstrak informasi penting dari data besar dan menganalisis struktur variabel. PCA dapat mengurangi dimensi data tinggi ke dimensi data rendah sekaligus menjaga risiko kehilangan informasi sangat rendah [9].

3.7. *Evaluation Metric*

Evaluatin metric adalah metode pengukuran yang digunakan untuk mengukur kualitas model Machine Learning yang dikembangkan. Metrik yang umum digunakan untuk mengukur kualitas model Machine Learning meliputi *accuracy, precision, recall, F1-score* dan *silhoutte score.*

3.7.1. *Accuracy*

Accuracy adalah salah satu indikator yang paling umum digunakan untuk mengukur kualitas model Machine Learning. Umumnya *accuracy* digunakan untuk mengukur rasio jumlah prediksi yang benar atas jumlah data uji keseluruhan [10]. Rumus dari *accuracy* dapat dilihat pada Rumus 1.

$$Accuracy = \frac{true\ positive + true\ negative}{total\ seluruh\ data} \quad (1)$$

Rumus 1 merupakan rumus dari *accuracy*. Rumus dari *accuracy* adalah menjumlahkan nilai true positive dengan true negative lalu membaginya dengan total seluruh data

3.7.2. *Precision*

Precision merupakan salah satu metode pengukuran yang berfungsi untuk mengukur jumlah yang diprediksi benar *(true positive)* dari keseluruhan data dalam kelas positif [10]. Rumus dari *precision* dapat dilihat pada Rumus 2.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2)$$

Rumus 2 merupakan rumus dari *precision*. Rumus dari *precision* adalah membagi nilai *true positive* dengan hasil penjumlahan *true positive* dan *false positive*.

3.7.3. *Recall*

Recall merupakan metode pengukuran model Machine Learning yang digunakan untuk mengukur rasio jumlah prediksi yang benar pada kelas positif atas keseluruhan data pada kelas positif [10]. Rumus dari *recall* dapat dilihat pada Rumus 3.

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (3)$$

Rumus 3 merupakan rumus dari *recall*. Rumus dari *recall* adalah nilai *true positive* dibagi dengan hasil dari penjumlahan nilai *true positive* dan *false negative.*

3.7.4. *F1-Score*

F1-score atau yang sering disebut juga dengan *F-measure* merupakan metode pengukuran model Machine Learning yang mewakili nilai rata-rata harmonik *(harmonic mean)* antara nilai *recall* dan *precision* [10]. Rumus dari *F1-score* dapat dilihat pada Rumus 4.

$$FM = 2\ x\ \frac{presisi\ x\ recall}{presisi + recall} \quad (4)$$

Rumus 4 merupakan rumus dari *F1-score.* Rumus *F1-score* adalah membagi hasil perkalian *precision* dan *recall* dengan hasil penjumlahan *precision* dan *recall*, lalu dikalikan dua.

3.7.5. *Silhouette Score*

Silhouette score atau yang sering disebut juga dengan *silhouette coefficient* merupakan metode pengukuran model Machine Learning yang mampu mengukur kualitas dan kekuatan *cluster*, sehingga dapat dilihat seberapa baik data ditempatkan dalam sebuah *cluster* [11]. Rumus dari *silhouette score* dapat dilihat pada Rumus 5.

$$Silhouette\ Score = \frac{(b-a)}{\max(a,b)} \quad (5)$$

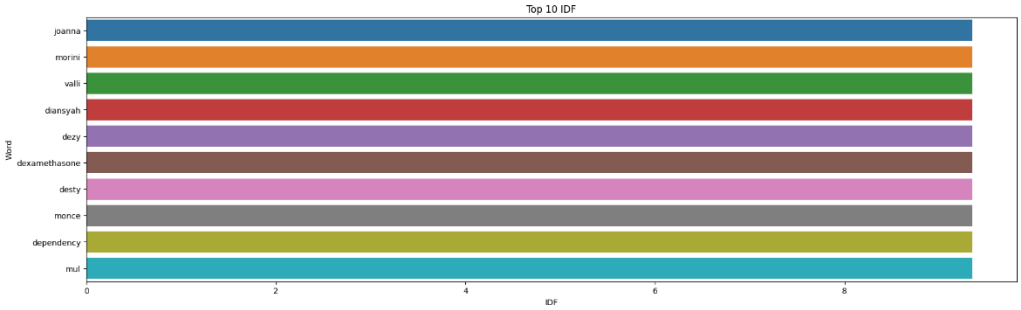
Rumus 5 merupakan rumus dari *silhouette score*. Nilai a merupakan jarak rata-rata *intra-cluster* dan nilai b merupakan jarak *mean-cluster* terdekat. Rumus dari *silhouette score* adalah membagi hasil pengurangan nilai jarak *mean-cluster* terdekat dan jarak rata-rata *intra-cluster* dengan nilai maksimum dari jarak rata-rata *intra-cluster* dan jarak *mean-cluster.*

4. Hasil dan Pembahasan

Penelitian yang dibuat melakukan pengujian dalam membandingkan nilai *cluster* pada K-Means. K-Means yang diterapkan pada penelitian ini menerapkan TF-IDF, *feature selection* dan PCA. Proses penilaian nilai *cluster* menggunakan visualisasi berupa *bar plot* dari tiap nilai metric yang diperhatikan yaitu *mean silhouette, accuracy, precision, recall, F1-score* dan *silhouette* score.

4.1. TF-IDF

TF-IDF dilakukan untuk melakukan proses ekstraksi fitur pada metode K-Means untuk pembobotan kata terhadap dataset yang digunakan. Hasil TF-IDF yang diterapkan pada metode K-Means dalam penelitian yang dibuat dijabarkan pada Gambar 2.

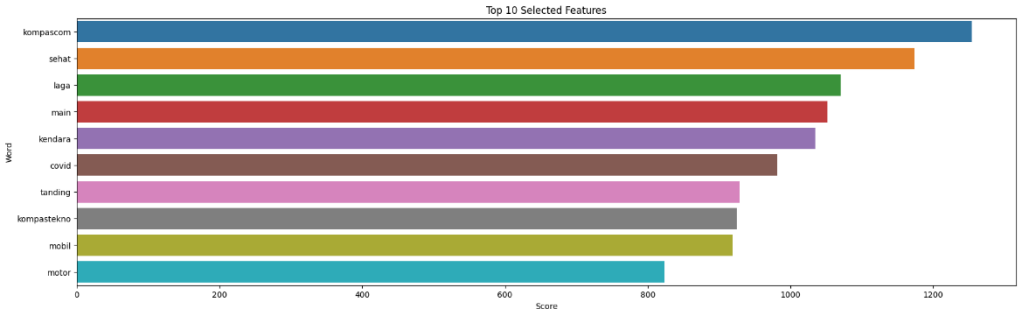


Gambar 2. Hasil TF-IDF K-Means

Gambar 2. merupakan hasil TF-IDF yang diterapkan pada metode K-Means dari penelitian ini. Hasil TF-IDF yang ditampilkan berupa 10 IDF tertinggi pada dataset yang digunakan, diantaranya joanna, morini, valli, diansyah, dezy, dexamethasone, desty, monce, depedency dan mul.

4.2. *Feature Selection*

Feature selection yang diterapkan pada penelitian yang dibuat adalah metode ANOVA. Hasil *feature selection* dari penelitian yang dibuat dapat dilihat pada Gambar 3.

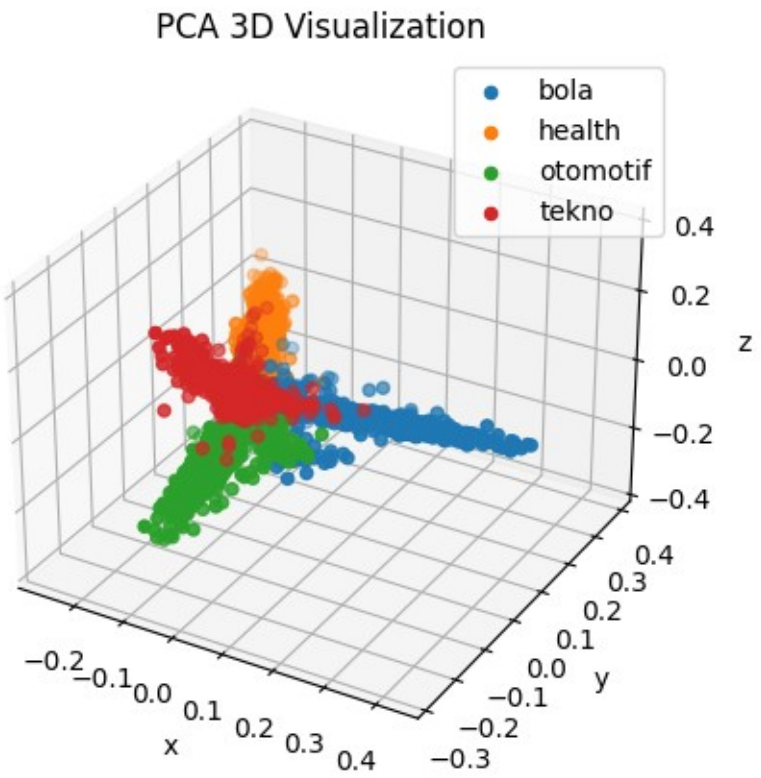


Gambar 3. Hasil *Feature Selection*

Gambar 3. merupakan hasil *feature selection* yang dilakukan dalam penelitian ini. Berdasarkan hasil *feature selection* pada Gambar 3, dapat dilihat 10 fitur tertinggi yang diterapkan dalam penelitian yang dibuat diantaranya kompascom, sehat, laga, main, kendara, covid, tanding, kompastekno, mobil dan motor.

4.3. PCA

Penelitian metode K-Means yang dilakukan menerapkan reduksi dimensi. PCA merupakan teknik reduksi dimensi yang diterapkan dalam penelitian yang dibuat. Hasil penerapan PCA dalam penelitian yang dibuat dapat dilihat pada Gambar 4.



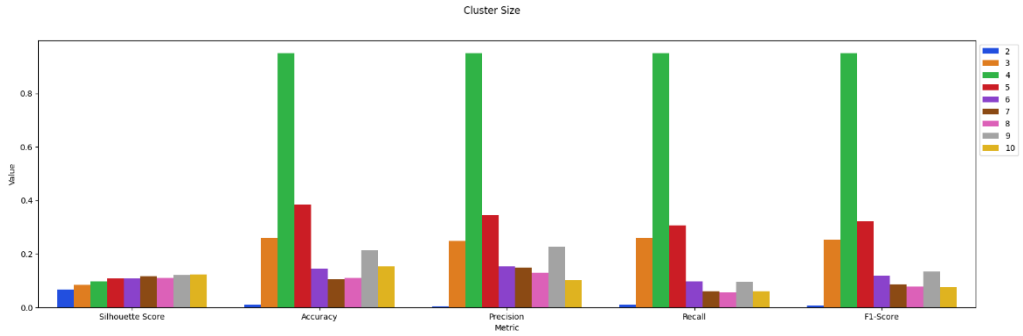
Gambar 4. Hasil PCA

Gambar 4. merupakan visualisasi PCA metode K-Means pada penelitian ini. Berdasarkan hasil visualisasi PCA pada Gambar 4., dapat dilihat masih ada sebaran data yang bertumpuk pada kategori lain. Bentuk data yang dihasilkan masih menyerupai bentuk oval/elips, meskipun masih terdapat sebaran data yang bertumpuk pada kategori lain. Metode K-Means jika data yang digunakan bentuknya berupa oval/elips maka hasilnya akan jauh lebih baik (optimal).

4.4. *Training*

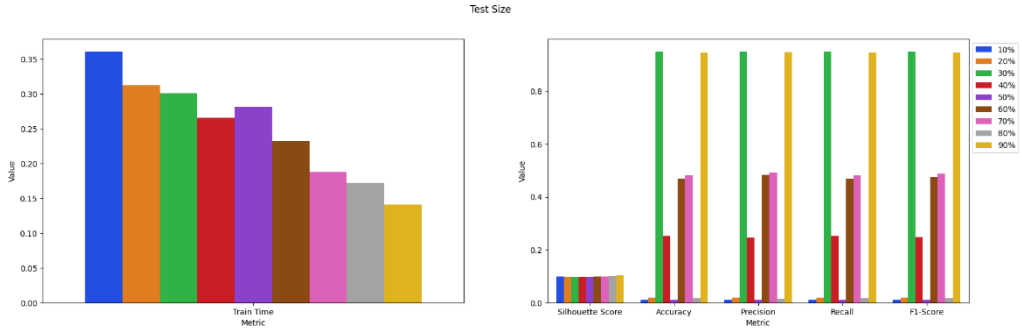
Training merupakan proses terakhir dari tahapan persiapan metode K-Means. Tahap ini bertujuan untuk melatih metode yang telah dikompilasi sebelumnya dengan menggunakan data latih yang telah dipersiapkan sehingga metode yang dibuat dapat dipergunakan untuk membuat suatu prediksi dari data yang baru.

Penelitian ini melakukan pengujian terhadap nilai *cluster* dan rasio data uji pada metode K-Means yang dibuat. Range nilai cluster yang diuji adalah 2 hingga 10. Rasio data uji 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 dan 0.9 merupakan rasio data uji yang digunakan pada penelitian yang dibuat. Hasil pengujian nilai *cluster* dan rasio data uji pada metode K-Means yang dibuat dapat dilihat pada Gambar 5. dan Gambar 6.



Gambar 5. Hasil Pengujian Nilai *Cluster*

Gambar 5. merupakan hasil pengujian nilai *cluster*, nilai *cluster* 4 memperoleh nilai terbaik pada *metric accuracy, precision, recall* dan *F1-score*. Nilai *cluster* 10 memiliki nilai terbaik pada *metric sillhouette score.* Nilai *cluster* 4 dipilih sebagai nilai *cluster* yang paling optimal dikarenakan mendominasi pada empat *metric*.



Gambar 6. Hasil Pengujian Rasio Data Uji K-Means

Gambar 6. merupakan hasil pengujian rasio data uji pada metode K-Means yang dibuat. Gambar 6. menunjukkan rasio data uji 30% memperoleh nilai terbaik disetiap metric kecuali training time, meskipun begitu rasio data uji 30% tetap dipilih sebagai rasio data uji yang paling optimal.

5. Kesimpulan

Kategori berita yang begitu banyak dan beragam mengakibatkan kinerja para editor semakin banyak. Proses pengelompokan berita secara manual memakan waktu lama dan tidak efisien. Kemajuan teknologi informasi belakangan ini dapat memberikan kemudahan untuk pekerjaan yang membutuhkan suatu proses pengelompokkan. Proses Clustering dapat digunakan untuk mengelompokkan data sehingga data pada kategori yang sama dapat dikelompokkan dengan data yang serupa atau sejenis. Penelitian yang dibuat menerpakan Metode K-Means. K-Means adalah teknik Clustering berbasis jarak yang dibagi menjadi serangkaian *cluster* dan hanya berfungsi untuk atribut numerik. Uji K-Means yang dilakukan dalam penelitian ini digunakan untuk membandingkan nilai *cluster*. Metode K-Means yang dibuat pada penelitian ini menerapkan TF-IDF, *feature selection* dan PCA. Proses penilaian nilai *cluster* menggunakan visualisasi berupa *bar plot* dari tiap nilai *metric* yang diperhatikan yaitu *mean silhouette, accuracy, precision, recall, F1-score* dan *silhouette score*. Hasil penelitian yang telah dilakukan Metode K-Means mampu mencapai 94.93% *accuracy* dan *recall,* 95.07% *precision* serta 94.94% *F1-score.*

References

[1] I. Setiawan and D. Nursantika, “Klasifikasi Artikel Berita Menggunakan Metode Text Mining Dan Naive Bayes Classifier,” *Pros. SENIATI*, pp. 1–6, 2017.

[2] M. Robani and A. Widodo, “Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia,” *J. Sist. Inf. Bisnis*, vol. 6, no. 2, p. 164,

2016.

[3] G. E. I. Kambey *et al.*, “Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia,” *J. Tek. Inform.*, vol. 15, no. 2, pp. 75–82, 2020.

[4] S. Gusriani, K. D. K. Wardhani, and M. I. Zul, “Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi Naïve Bayes (Studi Kasus: Facebook Page BerryBenka) Top Words Analysis of Online Media in Indonesia View project Wifi Positioning System (WPS) View project,” *Researchgate.Net*, no. September, 2016.

[5] A. S. M. Romli, *Jurnalistik Online: Panduan Mengelola Media Online*. Bandung: Nuansa Cendekia, 2018.

[6] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, “Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods,” *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 11, no. 1, p. 32, 2020.

[7] M. Z. Naf’an, A. Burhanuddin, and A. Riyani, “Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen,” *J. Linguist. Komputasional*, vol. 2, no. 1, pp. 23–27, 2019.

[8] U. Triharyuni, Setiya., Nugraha, Budi., Chodriyah, “Pengaruh Lama Setting Dan Jumlah Pancing Terhadap Hasil Tangkapan Rawai Tuna Di Laut Banda Influence of Setting Time and Numbers of Hooks At Tuna,” *J. Lit. Perikan. Ind*, vol. 19, pp. 81–88, 2013.

[9] A. S. Ritonga and I. Muhandhis, “Teknik Data Mining Untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data Principal Component Analysis (PCA),” *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 2, 2021.

[10] M. Hossin and M. N. Sulaiman, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015.

[11] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, “The K-Means Clustering Algorithm With Semantic Similarity To Estimate The Cost of Hospitalization,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 4, p. 313, 2019.