E-ISSN: 3031-8629
P-ISSN: 3032-3258
Halaman 194-207
Volume 1 Nomor 3 Tahun 2024

Klusterisasi Judul Berita Pada Website Detik Menggunakan Algoritma Kmeans

Adhif Adhitama[1], Syarif Hidayatullah[2] ,Miftahur Rahman[3]
[1,2,3] Program Studi Teknik Informatika, Universitas Muhammadiyah Jember
e-mail: adhifadhitama99@gmail.com syarifhidayatullah@unmuhjember.ac.id
miftahurrahman@unmuhjember.ac.id

Abstrak

Di era digital saat ini, pertumbuhan informasi yang cepat dan luas memunculkan tantangan dalam mengelola dan mengakses informasi yang relevan. Situs berita online seperti detik.com menyediakan platform yang penting bagi masyarakat untuk mendapatkan informasi terkini dan beragam. Dengan ratusan judul berita yang dipublikasikan setiap harinya, penting untuk dapat mengelompokkan berita-berita tersebut ke dalam kategori yang jelas dan mudah diakses. Klusterisasi judul berita adalah pendekatan yang dapat digunakan untuk mengelompokkan berita-berita tersebut berdasarkan kesamaan topik atau tema. Dalam konteks ini, sistem diharapkan mampu mengelompokkan berita sesuai dengan kesamaan yang dimiliki. Sehingga dari permasalahan tersebut penelitian ini akan menawarkan solusi yaitu melakukan Klusterisasi Judul Berita Pada Website Detik.com Menggunakan Kmeans untuk mengetahui pola kesamaan yang terdapat dijudul berita. Penentuan titik pusat awal (centroid) dalam Algoritma K-Means sangat mempengaruhi hasil klaster, karena centroid awal ditentukan secara acak. Hal ini dapat mengakibatkan hasil klaster yang kurang akurat. Oleh karena itu, diperlukan proses uji coba berulang untuk memperoleh hasil klaster yang optimal.

Kata kunci: *Klustering, Kmeans, Teks Mining, Berita Online*

Abstract

In today's digital era, the growth of rapid and widespread information creates challenges in managing and accessing relevant information. accessing relevant

information. Online news sites such as Detik.com provide an important platform for people to get up-to-date and diverse information. information. With hundreds of news titles published every day, it is important to be able to categorize the news into clear and accessible categories. that are clear and easily accessible. Headline clustering is an approach that approach that can be used to categorize the news based on similarity in topic or theme. In this context, the system is expected to be able to group news according to their similarities. So that from these problems This research will offer a solution, namely clustering news titles on the Detik website. Using Kmeans to find out the similarity patterns contained in news titles. Determination of the initial center point (centroid) in the K-Means Algorithm greatly affects the results of the cluster, because the centroid is very important. affects the cluster results, because the initial centroid is determined randomly. This This can result in less accurate cluster results. Therefore, therefore, an iterative trial process is required to obtain optimal cluster results. optimal cluster result.

Keywords: Clustering, Kmeans, Teks Mining, Online News

LATAR BELAKANG

Di era digital saat ini, pertumbuhan informasi yang cepat dan luas memunculkan tantangan dalam mengelola dan mengakses informasi yang relevan. Situs berita online seperti Detik.com (Sely Karmila, 2022) menyediakan platform yang penting bagi masyarakat untuk mendapatkan informasi terkini dan beragam. Dengan ratusan judul berita yang dipublikasikan setiap harinya, penting untuk dapat mengelompokkan berita-berita tersebut ke dalam kategori yang jelas dan mudah diakses. Klusterisasi judul berita adalah pendekatan yang dapat digunakan untuk mengelompokkan berita-berita tersebut berdasarkan kesamaan topik atau tema.

Clustering text (Wisna et al., 2023) melibatkan data dalam bentuk teks tidak terstruktur, di mana teknik dalam text mining digunakan untuk menemukan kata atau pola yang penting dalam proses clustering. Penambangan teks melibatkan identifikasi mendapatkan wawasan dan keahlian berharga melalui analisis pola di

berbagai sumber data teks. Proses penambangan teks seperti pelipatan kasus, tokenisasi, pemfilteran, dan stemming berupaya mengurangi atau mengekstraksi data dan meminimalkan gangguan dalam data. Diperlukan suatu metode untuk merepresentasikan data yang diproses dalam proses clustering sebagai vektor kata Memanfaatkan pembobotan kata (term Weighting) untuk menentukan seberapa sering setiap istilah muncul. Dalam situasi ini, teknik yang biasanya digunakan untuk pembobotan kata adalahTerm Frequency (Tf) dan Inverse Document Frequency (Idf) (Hanami, 2023).

Klasterisasi judul berita (Sely Karmila, 2022) merupakan teknik yang penting dalam analisis konten digital untuk mengelompokkan judul-judul berita yang serupa ke dalam kelompok-kelompok yang lebih terorganisir dan bermakna. Dalam konteks media digital, website berita seperti Detik merupakan sumber informasi yang menyajikan berbagai topik berita dari berbagai kategori, mulai dari politik, ekonomi, hingga hiburan, dengan frekuensi update yang tinggi.

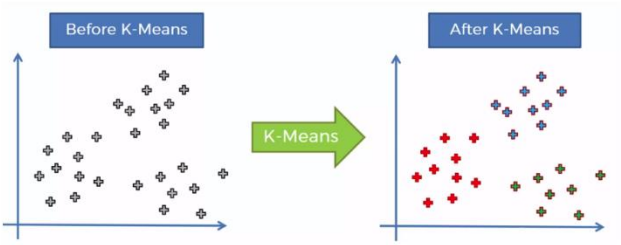
Alasan dipilihnya algoritma K-Means untuk penelitian ini karena data masukannya cukup mudah untuk diproses, K-Means dianggap lebih cocok untuk melakukan pengelompokan. Selain itu, penelitian sebelumnya menunjukkan bahwa Algoritma K-Means efektif dalam mengelompokkan dokumen teks. Misalnya, (Rachman et al., 2020) berhasil Dilakukan clustering skripsi mahasiswa Universitas Ma Chung dengan algoritma K-Means, menghasilkan nilai kemurnian sebesar 76%. Hal ini menunjukkan bahwa sistem telah mengkategorikan sekitar 76% dokumen secara akurat.

Dalam konteks ini, sistem diharapkan mampu mengelompokkan berita sesuai dengan kesamaan yang dimiliki. Sehingga dari permasalahan tersebut penelitian ini akan menawarkan solusi yaitu melakukan Klusterisasi Judul Berita Pada Website Detik Menggunakan Kmeans untuk mengetahui pola kesamaan yang terdapat dijudul berita.

E-ISSN: 3031-8629
P-ISSN: 3032-3258
Halaman 194-207
Volume 1 Nomor 3 Tahun 2024

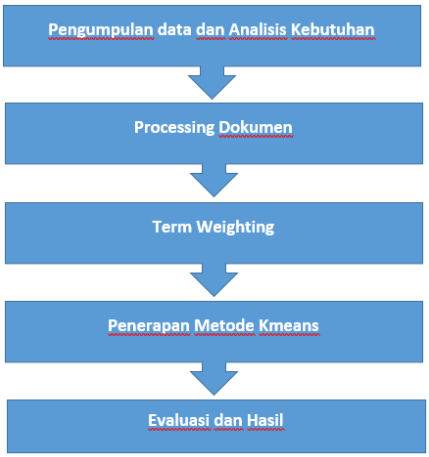
METODE

Clustering merupakan algoritma pengelompokkan sejumlah data menjadi kelompok–kelompok data tertentu (*cluster*), yang bertujuan untuk mengelompokkan data dengan beberapa karakteristik, baik dari segi wilayah ataupun kelompok yang berbeda, seperti Gambar dibawah ini:



Gambar 1. Contoh Pengelompokan Berdasarkan Bentuk dengan Kmeans

Data disusun berdasarkan atribut bentuknya, dikategorikan ke dalam kelompok persegi panjang, persegi, dan lingkaran untuk ditampilkan. Saat merancang sebuah penelitianan ini ada beberapa proses yang digunakan yaitu sebagai berikut:



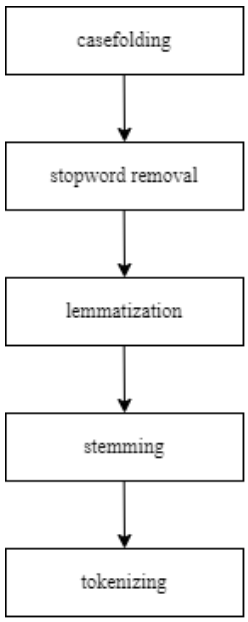
Gambar 2. Tahapan Penelitian

1. Pengumpulan Data dan Analisis Kebutuhan

Dalam pengumpulan peneliti mengambil dari data yang ada di Detik.com, adapun dataset yang tersedia yaitu dataset tahun 2020, dengan jumlah data yang didapat yaitu 1000 data, adapun yang akan penulis pakai hanya 100 data sebagai data sample.

2. Processing Dokumen

Pra-pemrosesan dokumen dilakukan untuk memudahkan proses clustering terhadap dokumen-dokumen yang diperoleh. Memulai pemrosesanberfungsi untuk meningkatkan citra (Krishnalal et al., 2010). Tahapan preprocessing teks (Kannan et al., 2014) Ini melibatkan serangkaian tindakan untuk mempersiapkan dan mengatur data teks untuk analisis tambahan. Langkah-langkah selanjutnya adalah fase umum dalam pra-pemrosesan. Berikut adalah alur dari proses preprocessing



Gambar 3. Alur Prepocessing Data

3. Term Weighting

Setelah prapemrosesan dokumen selesai (termasuk pelipatan kasus, tokenisasi, pemfilteran, dan stemming), serangkaian istilah atau kata akan dihasilkan oleh fase prapemrosesan. Ketentuan Frequency — Inverse Document Frequency (TF-IDF) adalah sebuah metode yang berguna untuk mengevaluasi bobot relatif dari kata-kata dalam sebuah dokumen atau kumpulan dokumen. Metode ini digunakan setelah tahap preprocessing data dan bertujuan untuk mengubah data teks yang tidak terstruktur menjadi lebih terstruktur (Hanami, 2023). Setelah proses pembobotan kata, data yang telah terstruktur dapat dikategorikan menggunakan berbagai metode klasifikasi. Nilai pembobotan kata dapat bervariasi tergantung pada teknik pembobotan yang digunakan. Metode-metode tersebut dapat mempengaruhi bagaimana kata-kata dinilai berdasarkan kepentingan dan kemunculannya di dalam dokumen atau korpus teks.

Dengan menggunakan TF-IDF, kata-kata yang sering muncul dalam sebuah dokumen tetapi jarang muncul dalam korpus secara keseluruhan akan memiliki bobot yang lebih tinggi, yang menunjukkan pentingnya kata-kata tersebut dalam dokumen tertentu. Ini membantu dalam identifikasi dan penekanan kata-kata kunci yang paling relevan untuk setiap dokumen atau kumpulan dokumen.

Metode pembobotan kata yang berbeda dapat menghasilkan nilai bobot yang beragam untuk kata-kata yang sama, tergantung pada metode spesifik yang digunakan dalam implementasinya. TF-IDF digunakan untuk mendapatkan nilai bobot kata pada 'Pertanyaan' dan 'Kategori', dan nilai bobot tersebut digunakan untuk mendapatkan nilai *cosine similarity*.

Dalam konteks Term Frequency (TF), terdapat beberapa formula yang digunakan untuk mengukur seberapa sering suatu kata muncul dalam sebuah dokumen. Nilai TF ini mencerminkan tingkat frekuensi kemunculan kata tersebut. Beberapa jenis formula yang umum digunakan melibatkan pendekatan berikut ini:

$$TF = \begin{cases} 1 + log\,10\ (f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases}$$

Nilai ft,d mewakili seberapa sering istilah (t) muncul dalam dokumen (d).

Oleh karena itu, jika suatu istilah ada dalam dokumen sebanyak 7 kali, maka bobot yang diberikan padanya adalah 1,845, dihitung 1 ditambah logaritma 7. Namun, jika istilah tersebut tidak ditemukan dalam dokumen, bobot yang diberikan padanya adalah nol. (0).

4. Proses Clustering

Algoritma *K-Means* dapat diterapkan dalam klasterisasi teks dengan langkah-langkah yang mirip dengan klasterisasi data pada umumnya, namun dengan penyesuaian pada representasi data teks.

Rumus-rumus

Jarak Euclidean antara dua titik x dan c_i

yang Digunakan:

$$d(x, c_i) = \sqrt{\sum_{j=1}^{n} (x_j - c_{ij})^2}$$

Dimana $x = x1, x2, \ldots, xn$ adalah titik data, $c_i = (c_{i1}, c_{i2,\ldots\ldots,} c_{in})$ adalah centroid dari cluster i dan n adalah dimensi dari ruang fitur.

Update Centroid:

Setelah menetapkan titik-titik ke dalam cluster, posisi baru dari centroid C_i dihitung sebagai rata-rata dari semua titik data dalam cluster tersebut: c

$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$$

Dimana S_i adalah himpunan semua titik data yang termasuk dalam cluster i

HASIL DAN PEMBAHASAN

Pembahasan

Tahapan pertama dalam penelitian ini adalah *preprocessing* dokumen, yaitu peneliti akan melakukan perhitungan terhadap 100 data percobaan, pada 1000 judul berita pada situs detik. Adapun bebeapa tahapan *preprocessing* dibawah ini adalah sample dari 100 data percobaan, data yang tertera adalah Date, Url, Judul, Kategori, sedangkan yang akan diproses adalah judul berita dan kategori. Adapun kategori yang diambil adalah secara acak, dengan rincian 10 judul berita finance, 4 judul berita food, 10 judul berita health, 5 judul berita hot, 20 judul berita internet, 20 Judul berita News, 15 judul berita Sport, 11 Judul berita Travel.

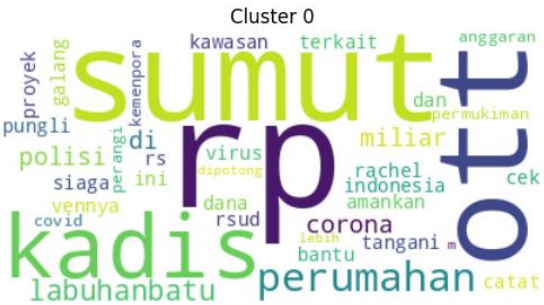
Hasil Cluster

Penentuan jenis konten dari setiap klaster dilakukan dengan menganalisis kata-kata yang paling sering muncul dalam masing-masing klaster. Kata yang paling sering muncul akan diambil untuk menggambarkan jenis dari judul berita pada setiap klaster

Tabel 1 Hasil Cluster 1

Kata sering Muncul	kategori
rachel vennya galang dana rp 1 miliar bantu tangani corona di indonesia	health
catat! ini 5 rs siaga virus corona di sumut	news
ott kadis perumahan labuhanbatu sumut, polisi amankan cek rp 1,4 miliar	news
ott kadis perumahan labuhanbatu sumut terkait pungli proyek rsud	news

Kata sering Muncul	kategori
polisi ott kadis perumahan dan kawasan permukiman labuhanbatu	news
perangi covid-19, anggaran kemenpora dipotong rp 500 m lebih	finance



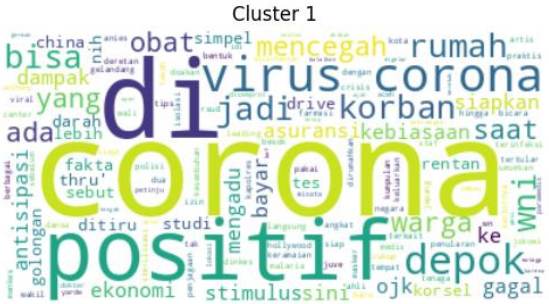
Gambar 4 adalah visualisasi dari cluster 1

Tabel 2 Hasil Cluster 2

Kata sering Muncul	kategori
ojk siapkan stimulus ekonomi antisipasi dampak corona	finance
jadi korban gagal bayar asuransi? bisa mengadu ke sini	finance
5 kebiasaan simpel yang bisa mencegah virus corona	health
bisa ditiru nih! 5 fakta tes corona 'drive-thru' di korsel	health
studi di china sebut golongan darah a lebih rentan terinfeksi corona	health
dua staf who positif tertular virus corona	health

E-ISSN: 3031-8629
P-ISSN: 3032-3258

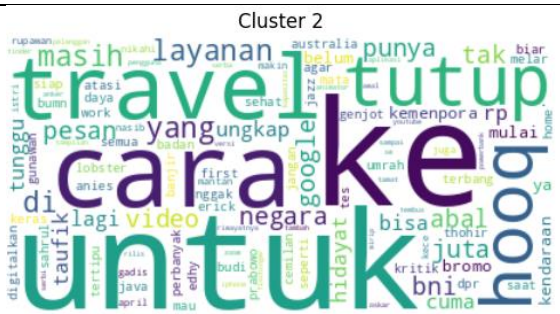
Halaman 194-207
Volume 1 Nomor 3 Tahun 2024



Gambar 5 *Cluster* 2

Tabel 3 Hasil Cluster 3

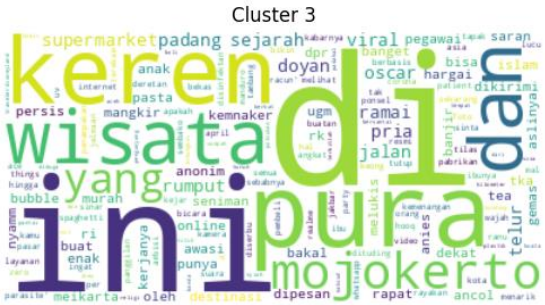
Kata sering Muncul	kategori
bni digitalkan bni java jazz 2020	finance
terbang ke australia, edhy prabowo mau genjot budi daya lobster	finance
cara biar nggak tertipu travel umrah abal-abal seperti first travel	finance
dpr kritik cara anies atasi banjir, pesan keras erick thohir ke BUMN	finance
pesan who untuk semua negara: perbanyak tes, jangan tutup mata!	health
5 cemilan sehat agar badan tak melar saat 'work from home'	health



Gambar 6 *Cluster* 3

Tabel 4 Hasil Cluster 4

Kata sering Muncul	kategori
kemnaker awasi tka di meikarta	finance
saran buat anies-rk yang mangkir rapat banjir di dpr	finance
ri bakal punya destinasi wisata sejarah islam di ancol	finance
pegawai supermarket ini dikirimi bubble tea oleh anonim yang hargai kerjanya	food
doyan banget telur, seniman ini melukis telur persis aslinya	food
nyamm! pasta murah dan enak dekat ugm ini bisa dipesan online	food



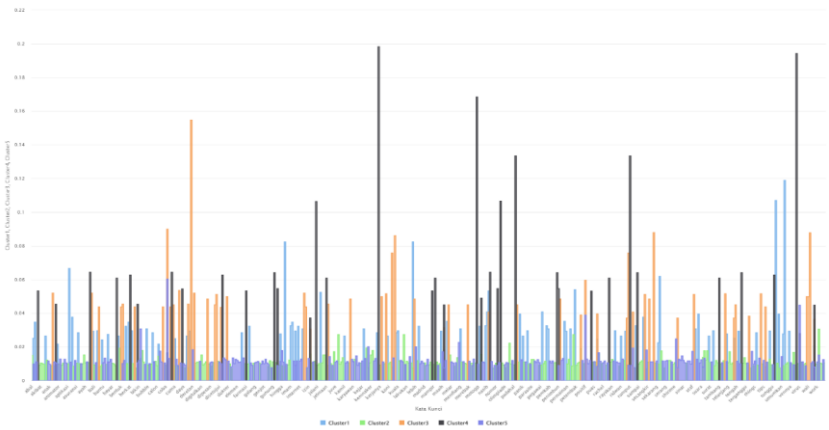
Gambar 7 *Cluster* 4

Tabel 5 Hasil Cluster 5

Kata sering Muncul	kategori
beda 19 tahun, sahrul gunawan sebut calon istri sudah dekat dengan anaknya	hot
meme sindir penimbun masker, kecele harga sudah balik normal	internet
viral cuitan harga masker mulai normal, netizen bahas penimbun	internet
ridwan kamil: jangan panik beli masker-sembako, supply demand kita aman	news
tiket olimpiade sudah, petembak vidya rafika akan jalani uji coba	sport



Gambar 8 *Cluster* 5



Gambar 9 Grafik Rata-Rata Jumlah yang sering keluar dalam cluster judul berita Masing-masing cluster memiliki judul dengan jumlah kata yang paling sering muncul

KESIMPULAN

Temuan penelitian menunjukkan bahwa pemilihan centroid awal pada algoritma K-Means mempunyai dampak yang signifikan terhadap hasil cluster karena proses pemilihannya secara acak. Hal ini dapat mengakibatkan hasil klaster yang kurang akurat. Oleh karena itu, diperlukan proses uji coba berulang untuk memperoleh hasil klaster yang optimal.

DAFTAR PUSTAKA

Afida, D., Udayanti, E. D., & Kartikadarma, E. (2021). Aplikasi Text Mining untuk Klasterisasi Aduan Masyarakat Kota Semarang Menggunakan Algoritma K-means. *Jurnal Transformatika*, *18*(2), 215. https://doi.org/10.26623/transformatika.v18i2.2362

Hanami, A. (2023). Term Frequency Inverse Document Frequency (TF-IDF). *Visitor Analytics*, *December*. https://www.visitor-analytics.io/es/glosario/t/term-frequency-inverse-document-frequency-tf-idf

Herlawati, H., & Handayanto, R. T. (2020). Penggunaan Matlab dan Python dalam Klasterisasi Data. *Jurnal Kajian Ilmiah*, *20*(1), 103–118.

https://doi.org/10.31599/jki.v20i1.85

Kannan, S., Gurusamy, V., & Kannan S. (2014). Preprocessing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Krishnalal, G., Rengarajan, S. B., & Srinivasagan, K. G. (2010). A New Text Mining Approach Based on HMM-SVM for Web News Classification. *International Journal of Computer Applications*, *1*(19), 103–109. https://doi.org/10.5120/395-589

Rachman, D. A. C., Goejantoro, R., & Amijaya, F. D. T. (2020). Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering Implementation Of Text Mining For Grouping Thesis Documents Using K-Means Clustering. *Jurnal EKSPONENSIAL*, *11*(2), 167–174.

Satria, G. J., Adikara, P. P., & Wihandika, R. C. (2022). Klasifikasi Pertanyaan COVID-19 Bahasa Indonesia menggunakan Naive Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(1), 148–153. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10382

Sely Karmila, V. I. A. (2022). METODE LATENT DIRICHLET ALLOCATION UNTUK MENENTUKAN TOPIK TEKS SUATU BERITANo Title. *Jurnal Informatika & Komputasi*, *16*.

Siringoringo, R., & Jamaludin, J. (2019). Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online. *Jurnal Teknologi Dan Ilmu Komputer Prima (JUTIKOMP)*, *2*(1), 41–48. https://doi.org/10.34012/jutikomp.v2i1.456

Tampubolon, P. P., Kaloka, T. P., Swasti, O., Mustika, W. F., & Bustamam, A. (2018). Model Klastering Skm3 (Subcontrolled K-Means Max-Min) Dan Aplikasinya Dalam Menghitung Elektabilitas Pasangan Calon Kepala Daerah. *Journal of Mathematics and Mathematics Education*, *8*(2), 85–96. https://doi.org/10.20961/jmme.v8i2.25838