CLUSTERING BERITA SEPAK BOLA DENGAN METODE K-MEANS

Radika Yudha Riyanto [1)], Viny Christanti Mawardi [2)], Novario Jaya Perdana [3)]

[1) 2) 3)] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
radika.535190028@stu.untar.ac.id [1)] vinym@fti.untar.ac.id [2)] novariojp@fti.untar.ac.id [3)]

ABSTRACT

Until now, many Indonesian people like soccer, both domestically and abroad. With so many football enthusiasts, people are becoming more active in finding news related to football. As time goes by, the amount of news circulating on the internet will also be more and more widespread. The large number of news makes the news need to be clustered or clustered to make it easier to access existing news. The website created is intended to group soccer news from several websites, namely: vivagoal.com, goal.com and bolasport.com. The method used on this website is K-Means to group news into clusters, then the method used to evaluate the quality of the clusters formed is the Silhouette coefficient method. The Silhouette coefficient value is 0.54, which means that the quality of the cluster formed is moderate.

***Keywords**: K-Means; Football; Clusters; News;*

1. Pendahuluan

Di era digital saat ini teknologi berkembang dengan sangat pesat, banyak keuntungan yang didapatkan dari berkembangnya teknologi. Dengan berkembangnya teknologi, informasi sangat mudah didapatkan dari berbagai media salah satunya adalah internet. Seiring dengan perkembangan internet, pengaruhnya terhadap masyarakat juga sangat penting, salah satunya untuk mempermudah pengumpulan informasi tentang isu-isu terkini, salah satunya melalui jejaring sosial. [1] Kebiasaan masyarakat juga ikut berubah dengan adanya perkembangan teknologi, untuk mendapatkan informasi masyarakat dapat dengan mudah mengaksesnya melalui media sosial, *website* ataupun sumber informasi lainnya yang dapat terhubung melalui internet.

Seiring berjalannya waktu, jumlah berita yang beredar di internet juga akan semakin banyak dan semakin luas. Banyaknya jumlah berita membuat berita tersebut perlu dilakukan klasterisasi atau *clustering* agar memudahkan pengguna dalam mengakses berita yang ada. *Clustering* dapat dipahami sebagai salah satu teknik *text mining* yang digunakan untuk pengelompokan dokumen, di mana dokumen dikelompokkan bersama dengan konten seperti berita tanpa identifikasi kategori sebelumnya. [2] Sehingga dibutuhkan satu aplikasi yang bisa mengelompokan sesuai dari judul berita yang ada, metode yang digunakan adalah *clustering*, *clustering* memiliki macam-macam algoritma salah satunya adalah K-Means karena algoritma K-Means cukup popular dan cukup mudah untuk diimplementasikan. Berita sepak bola yang beredar di internet sudah sangat banyak, berkaitan dengan hal tersebut kelebihan dari metode K-Means dapat sangat berguna untuk mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Kelemahan dari algoritma K-Means adalah sensitif terhadap inisialisasi *cluster*. [3]

Sistem ini dirancang untuk mengambil data dari berita yang terdapat pada *website* seperti www.goal.com, www.bola.net, dan www.vivagoal.com untuk melakukan training pada proses *Clustering*. Program ini dijalankan dengan melakukan input berupa dokumen berita dari *website* yang telah disebtukan. Ada empat tahapan yang akan dilalui dalam pembuatan sistem *Clustering* berita, yaitu preprocessing, term weighting, *Clustering* dengan menggunakan K-Means, dan evaluasi dengan mengguakan *Silhouette coefficient*. Hasilnya akan ditampilkan berupa *cluster-cluster* yang di dalam setiap *cluster* tersebut terdapat dokumen berita yang berasal dari *website* yang sudah ditentukan dan juga sudah dikelompokkan. Pembuatan rancangan akan menggunakan framework Flask untuk menampilkan tampilan pada *website*. HTML dan CSS digunakan untuk mengatur tampilan pada *website* agar dapat berinteraksi dengan *user*, Python akan digunakan untuk melakukan perhitungan dengan metode K-Means serta menggunakan metode *Silhouette coefficient* untuk evaluasi. Data akan diambil dengan cara melakukan scrapping dari tiga *website* berita sepak bola, kemudian data tersebut dilakukan preprocessing, tf-idf, dan K-Means. Dari *cluster* yang terbentuk akan dilakukan evaluasi dengan menggunakan *Silhouette coefficient*.

2. Landasan Teori

Proses pada rancangan ini akan dimulai dengan melakukan pengumpulan data berupa berita dari websiten berbahasa indonesia seperti www.goal.com,

www.bola.net, dan www.vivagoal.com. Pengumpulan data dilakukan dengan cara scraping dari website yang menyediakan berita mengenai sepak bola. Kemudian dilakukan pre-processing yang terdiri dari *case folding*, *tokenizing*, *filtering*, dan *stemming*. Setelah dilakukan pre-processing kemudian dilanjutkan dengan pembobotan TF-IDF agar data tersebut siap dilakukan clustering. Data yang sudah dilakukan pembersihan akan dihitung menggunakan metode K-Means Clustering. Kemudian untuk menghitung akurasi dari setiap cluster menggunakan *silhouette coefficient*.

Pembuatan rancangan akan menggunakan framework Flask untuk menampilkan tampilan pada website. HTML dan CSS digunakan untuk mengatur tampilan pada website agar dapat berinteraksi dengan user, Python akan digunakan untuk melakukan perhitungan dengan metode K-Means serta menggunakan metode *Silhouette coefficient* untuk evaluasi.

2.1. Berita Sepak Bola

Berita adalah fakta atau opini yang ingin diketahui banyak orang. Berita dapat diperoleh melalui berbagai media seperti surat kabar, surat kabar, televisi, internet dan lain-lain. [2] Berita biasanya bisa berbentuk lisan maupun tulisan, berita yang disampaikan secara lisan biasanya banyak ditemukan pada radio, televisi, maupun media sosial seperti youtube, dan tiktok sedangkan untuk penyampaian secara tertulis biasanya dapat berupa media cetak seperti koran, majalan ataupun media online seperti website dan media sosial lainnya.

Sepak bola adalah olahraga yang dimainkan oleh dua tim yang masing-masing terdiri dari 11 pemain. Tujuan utama dari permainan sepak bola adalah agar setiap tim berusaha untuk mendapatkan bola atau mencetak gol sebanyak-banyaknya ke gawang lawan dan melindungi gawangnya sendiri agar tidak kebobolan. Sebuah tim dinyatakan sebagai pemenang jika tim tersebut memasukkan bola terbanyak ke gawang lawan dan jika bola yang dimasukkan berjumlah sama, pertandingan dinyatakan seri. [4]

2.2. Text Mining

Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dimana, *text mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. [5]

1. *Case folding*

Case folding merupakan tahap text processing atau pra-pemrosesan yang mengubah semua huruf pada kalimat jika terdapat huruf besar menjadi huruf kecil yang bertujuan agar semua huruf yang akan diproses menjadi setara atau sama.

2. *Tokenizing*

Tokenizing merupakan tahap pemotongan kalimat menjadi per satu kata yang menyusun kalimat tersebut. Tiap potongan kata tersebut disebut sebagai token. Pemisahan kata tersebut dilakukan berdasarkan spasi yang berada dalam suatu kalimat.

3. *Filtering*

Filtering merupakan tahapan yang dilakukan untuk menghapus kata-kata yang kurang penting yang sering muncul di dalam sebuah teks dokumen. Tahap ini dilakukan untuk mempermudah proses perhitungan. Pada tahap ini menggunakan algoritma stoplist untuk membuang kata-kata yang kurang penting.

4. *Stemming*

Stemming adalah tahap mencari kata dasar dari setiap kata yang dihasilkan pada tahap *filtering* sebelumnya. Pada tahap ini dilakukan pemotongan atau penghapusan imbuhan. [3]

2.3. TF-IDF (*Term Frequency – Inverse Document Frequency*)

Term frequency atau TF merupakan menghitung bobot kata dengan menjumlahkan kata yang muncul pada dokumen tersebut. Sedangkan *Inverse Document Frequency* atau IDF merupakan jumlah kemunculan suatu kata pada semua dokumen yang ada. rumus perhitungan *Term frequency* (TF) dapat dilihat pada persamaan 1.

$$tf_{td} = f_{td} \quad (1)$$

Keterangan:

tf_{td} = Nilai term frequency

f_{td} = Frekuensi dari term pada data d

Rumus perhitungan Inverse Document Frequency (IDF) dapat dilihat pada persamaan 2.

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (2)$$

Keterangan:

idf_t = nilai IDF dari term t

N = Banyak Dokumen

df_t = Nilai DF dari term t

Rumus untuk menghitung term weighting TF-IDF yang merupakan gabungan dari TF dan IDF dapat dilhat pada persamaan 3. [6]

$$w_{t,d} = tf_{t,d} \times idf_t \quad (3)$$

Keterangan:

$w_{t,d}$ = Nilai Bobot dari t (term) dalam satu dokumen

$tf_{t,d}$ = frekuensi kemunculan t (term) pada dokumen d

idf_t = Nilai dari idf dari term t

2.4. K-Means Clustering

K-Means *Clustering* merupakan salah satu metode *cluster analysis non hirarki* yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam *cluster* yang lain. [7]

Dalam melakukan *Clustering* dengan metode K-Means ada beberapa tahapan yang harus dilakukan antara lain sebagai berikut:

1. Menentukan nilai k. Biasanya dilakukan dengan menentukan nilai random, namun pada rancangan ini untuk menentukan nilai k dapat menggunakan rumus pada persamaan 4. [3]

$$k = \sqrt{n/2} \quad (4)$$

Dimana n adalah jumlah total dokumen.

2. Tentukan titik pusat *cluster* secara acak yang berasal dari data yang akan diproses.
3. Hitung jarak antara data dan pusat *cluster* dengan menggunakan Manhattan Distance. Untuk menghitung jarak semua data terhadap titik pusat *cluster* dapat menggunakan rumus Manhattan Distance pada persamaan 5. [8]

$$d_{ij} = \sum_{i=1}^{n} |x_i - y_i| \quad (5)$$

Keterangan:

dij = Jarak data i ke pusat *cluster* j

xi = nilai objek i pada variabel x

yi = nilai objek i pada variabel y

n = banyaknya variabel yang diamati

4. Perbaharui nilai titik tengah *cluster* dengan cara mencari rata-rata data. Rata-rata jarak antar data dapat dilihat pada persamaan 6.

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i \quad (6)$$

Keterangan:

μk = Titik pusat *cluster* baru

Nk = Jumlah data yang terdapat pada *cluster* k

xi = Data yang terdapat pada *cluster* k

5. Ulangi langkah ketiga dan keempat hingga titik pusat *cluster* pada semua kelompok tidak lagi berubah. [3]

2.5. Silhouette coefficient

Metode *Silhouette coefficient* merupakan gabungan dari metode cohesion dan separation. Metode ini sering digunakan untuk melihat kualitas dan kekuatan *cluster* yaitu seberapa baik suatu objek ditempatkan dalam suatu *cluster*. [6] Metode cohesion digunakan untuk mengukur seberapa dekat relasi antara objek dalam sebuah *cluster*. Sedangkan metode separation yang berfungsi untuk mengukur seberapa jauh sebuah *cluster* terpisah dengan *cluster* lain. [9] Tahapan perhitungan Silhoutte coefficient adalah sebagai berikut:

1. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu *cluster* dengan rumus pada persamaan 7.

$$a(i) = \frac{1}{[A]-1} \sum j \in_{A, j \neq i} d(i,j) \quad (7)$$

2. Hitung rata-rata jarak dari objek ke-i tersebut dengan semua objek pada *cluster* lainnya, kemudian ambillah nilai terkecilnya dengan rumus pada persamaan 8.

$$d(i,C) = \frac{1}{[C]} \sum j \in C \quad d(i,j) \quad (8)$$

3. Setelah menghitung d(i, C) untuk semua C, cari nilai minimum dengan menggunakan rumus pada persamaan 9.

$$b(i) = \min C \neq A \quad d(i,C) \quad (9)$$

4. Nilai *silhouette coefficient* dapat dirumuskan seperti pada persamaan 10.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (10)$$

Nilai yang didapat dari metode *silhouette coefficient* terletak pada kisaran nilai -1 hingga 1. Jika nilainya mendekati 1, maka semakin baik pengelompokkan objeknya dalam satu *cluster*. Sebaliknya jika *silhouette coefficient* mendekati -1, maka semakin buruk pengelompokkan objeknya didalam satu *cluster*. [10] Nilai dan struktur *silhouette coefficient* dapat dilihat pada **Tabel 1.**

Tabel 1. Nilai *Silhouette coefficient*

Nilai *Silhouette coefficient*	**Struktur**
0.7 < *Silhouette coefficient* <= 1	Kuat
0.5 < *Silhouette coefficient* <= 0.7	Sedang
0.25 < *Silhouette coefficient* <= 0.5	Lemah
Silhouette coefficient <= 0.25	Tidak Terstruktur

3. Hasil Pengujian

3.1. Tampilan

1. Modul Beranda

Model beranda merupakan halaman yang pertama kali pengguna lihat ketika masuk ke dalam website clustering berita sepak bola. Pada halaman tersebut berisi kata sambutan, informasi singkat mengenai website dan 1 buah tombol, yaitu tombol "Clustering". Tombol tersebut berfungsi untuk melakukan navigasi ke halaman clustering. Modul beranda dapat dilihat pada **Gambar 1**.

Gambar 1. Modul Home

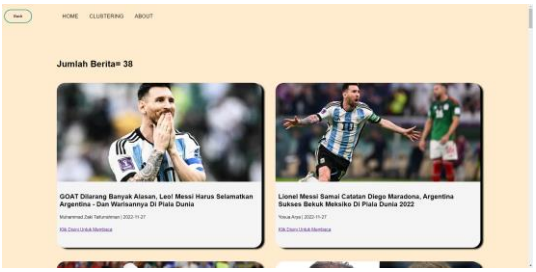
2. Modul Clustering

Modul Clustering merupakan modul dimana pengguna dapat melihat list cluster yang dimiliki website clustering berita sepak bola. Pada modul ini pengguna dapat melakukan klik pada salah satu button cluster yang tersedia dan juga pengguna dapat melihat akurasi dari cluster yang terbentuk dengan melihat nilai silhouette ceofficient. Modul Clustering dapat dilihat pada **Gambar 2**.

Gambar 2. Modul Clustering

3. Modul Berita

Modul Berita berisi list berita yang terdapat pada salah satu cluster yang dipilih di halaman clustering. Pada modul ini pengguna dapat melihat list berita yang terkelompokkan dalam suatu cluster, dan jika ingin membaca berita tersebut dengan lebih detail pengguna dapat melakukan klik pada tulisan "klik untuk membaca". Modul berita dapat dilihat pada **Gambar 3**.



Gambar 3. Modul Berita

4. Modul About

Modul about berisi informasi pembuat website dan juga fungsi dari website ini. Pada modul ini pengguna dapat melihat informasi dari pembuat website dan juga deskripsi mengenai website. Modul about dapat dilihat pada **Gambar 4**.

Gambar 4. Modul About

3.2. Pengujian Sistem

Pengujian terhadap sistem merupakan tahap pengujian untuk melihat evaluasi dari *clustering* dengan metode K-Means. Pengujian yang dilakukan menggunakan metode *silhouette coefficient* untuk menguji kualitas dari *cluster* yang terbentuk dari 597 data berita sepak bola dari 3 *website* yang telah

ditentukan. Hasil evaluasi dengan *silhouette coefficient* dapat dilihat pada **Tabel 2**.

Tabel 2. Hasil *Silhouette coefficient*

N_Clusters	*Silhouette coefficient*
2	0.6240529700538202
3	0.5857565845444187
4	0.5645811310823329
5	0.5524037756017318
6	0.5446137014943514
7	0.5357699807024148
8	0.5274095548450872
9	0.5247327422026177
10	0.5256749010368843
11	0.5190249836169671
12	0.5195473007361535
13	0.5118784843138234
14	0.5177790776494136
15	0.5239102890365747
16	0.5356948906847913
17	0.5432720989024539

Hasil Pengujian dari **Tabel 2**, dapat dilihat bahwa semakin banyak *cluster* yang terbentuk maka nilai *silhouette coefficient* nya semakin kecil. Hal ini dapat terjadi karena data yang di proses kurang banyak atau hanya sedikit memiliki kemiripan antar data sehingga saat masuk kedalam suatu *cluster* memiliki hasil yang kurang akurat. *Silhouette coefficient* juga hanya salah satu metode untuk melakukan evaluasi *cluster* dan memungkinkan bahwa metode evaluasi ini tidak selalu tepat untuk semua jenis kumpulan data.

Dari 597 data yang digunakan diambil nilai k dengan rumus **persamaan 1** yang menghasilkan banyak *cluster* yang terbentuk adalah 17 dan kemudian percobaan yang dilakukan untuk menguji kualitas *cluster* yang terbentuk mendapatkan nilai 0.5441564416964738 yang mana angka ini dapat diartikan bahwa kualitas dari *cluster* yang terbentuk tergolong sedang berdasarkan standar *silhouette coefficient* yang terdapat pada **tabel 1.**

4. Kesimpulan dan Saran

4.1. Kesimpulan

Berdasarkan hasil dan pembahasan pengujian yang telah dilakukan dalam *clustering* berita sepak bola dengan metode K-Means didapat kesimpulan sebagai berikut:

1. Algoritma K-Means dapat digunakan untuk melakukan clustering berita sepak bola dengan jumlah cluster adalah 17 hasil akurasi cluster yang dihitung dengan nilai *silhouette coefficient* mendapatkan nilai sebesar 0.5441564416964738.
2. Dari data yang diproses, semakin banyak cluster yang terbentuk maka nilai *silhouette coefficient* nya semakin kecil. Hal ini dapat terjadi karena data yang diproses hanya sedikit memiliki kemiripan antar data sehingga saat masuk kedalam suatu cluster memiliki hasil yang kurang akurat. *Silhouette coefficient* juga hanya salah satu metrik untuk melakukan evaluasi cluster dan memungkinkan bahwa metode evaluasi ini tidak selalu tepat untuk semua jenis kumpulan data.

4.2. Saran

Berdasarkan hasil dan pembahasan pengujian yang telah dilakukan dalam *clustering* berita sepak bola dengan metode K-Means didapat kesimpulan sebagai berikut:

1. Akurasi *silhouette coefficient* yang masih belum sempurna dapat diperbaiki dengan memperbaiki teknik pre-processing.
2. Data berita perlu di tambahkan agar akurasi dari *silhouette coefficient* semakin baik.
3. Perlu dilakukan percobaan untuk menentukan centroid yang baik dan cocok dengan data.
4. Untuk penelitian selanjutnya dapat menggunakan metode clustering lain selain K-Means Clustering dengan data/studi kasus yang sama.

REFERENSI

[1] Anita Dan J. Oliando, "Pengelompokkan Berita Kesehatan Pada Sosial Media Twitter Dengan Metode K-Means Clustering," Jurnal Ensiklopedia, Vol. 2, No. 3, Pp. 116-124, 2022.

[2] A. Y. Rofiqi, "Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat," Jurnal Simantec,Vol. 6, No. 1 2017.

[3] E. Susanto, V. C. Mawardi Dan M. D. Lauro, "Aplikasi Clustering Berita Dengan Metode K Means Dan Peringkas Berita Dengan Metode Maximum Marginal Relevance," Jurnal Ilmu Komputer Dan Sistem Informasi Aplikasi,Vol. 9, No. 1, Pp. 62-68, 2021.

[4] A. S. Nosa, "Survei Tingkat Kebugaran Jasmani Pada Pemain Persatuan Sepakbola Indonesia Lumajang," Jurnal Prestasi Olahraga, Vol. 1, No. 1, 2013.

[5] Y. S. Dan B. Wasito, "Analisis Testimonial Wisatawan Menggunakan Text Mining Dengan Metode Naive Bayes Dan Decision Tree, Studi Kasus Pada Hotel - Hotel Di Jakarta," Jurnal Informatika Dan Bisnis, Vol. 3, No. 2, Pp. 39-49, 2014.

[6] R. Handoyo, R. R. M Dan S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan Dokumen," Jurnal Sifo Mikroskil, Vol. 15, No. 2, Pp. 73-82, 2014.

[7] A. N. Khomarudin, "Teknik Data Mining : Algoritma K-Means Clustering," Jurnal Ilmu Komputer, Pp. 1-12, 2016.

[8] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, Dan Manhattan Distance Pada

Algoritma K- Means Clustering Berbasis Chi-Square," Jurnal Informatika: Jurnal Pengembangan IT (Jpit), Vol. 4, No. 01, Pp. 20-24, 2019.

[9] B. Wira, A. E. Budianto Dan A. S. Wiguna, "Implementasi Metode K-Medoids Clustering Untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 Di Universitas Kanjuruhan Malang," Jurnal Terapan Sains & Teknologi,Vol. 1, No. 3, Pp. 53-68, 2019.

[10] M. A. Nahdliyah, T. Widiharih Dan A. Prahutama, "Metode K-Medoids Clustering Dengan Validasi Silhouette Index Dan C-Index," Jurnal Gaussian,Vol. 8, No. 2, Pp. 161-170, 2019.

Radika Yudha Riyanto, seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara.

Viny Christanti Mawardi, memperoleh gelar S.Kom dari Universitas Tarumanagara tahun 2004 dan M.Kom dari Universitas Indonesia tahun 2008. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika, Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Novario Jaya Perdana, memperoleh gelar S.Kom dari ITS tahun 2011 dan M.T. dari Universitas Indonesia tahun 2016. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta