APLIKASI *CLUSTERING* BERITA DENGAN METODE K *MEANS* DAN PERINGKAS BERITA DENGAN METODE *MAXIMUM MARGINAL RELEVANCE*

Edy Susanto [1], Viny Christanti Mawardi [2], Manatap Dolok Lauro [3]

[1] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
edy.535170077@stu.untar.ac.id [1] , vinym@fti.untar.ac.id [2], manataps@fti.untar.ac.id [3]

ABSTRACT

News is information about facts or opinions that are interesting to know. News can be obtained from various media such as newspapers and the internet. As is well known, news has various topics, such as politics, sports and others. There is also the same story written with the addition of a little information. This causes it to take more time to get the headline of the news. Therefore we need a system for news clustering using the K-Means method and news summarizing using the Maximum Marginal Relevance (MMR) method in order to obtain information from news more easily and efficiently. News that is processed in the form of a collection of files (multi document) with the extension txt. The summarization process goes through the text preprocessing stage, which consists of sentence segmentation, case folding, tokenizing, filtering, stemming. The next step is TF-IDF calculation to calculate word weight then Cosine Similarity to calculate the similarity between documents. After that, enter the K-Means stage for clustering division and proceed with determining the summary with MMR. Based on the results testing that has been done, this application is running well, the results of clustering and summarizing news can make it easier for users to get news summaries from some similar news.

Key words

Berita, Clustering, K-Means, Maximum Marginal Relevance, Peringkasan teks.

1. Pendahuluan

1.1 Latar Belakang

Berita adalah sebuah informasi mengenai fakta atau opini yang menarik orang untuk diketahui. Berita bisa didapatkan dari berbagai media seperti surat kabar dan internet. [1]

Seperti yang diketahui, berita memiliki berbagai macam topik, seperti tentang kejahatan, politik, olahraga dan lain-lain. Dengan demikian, untuk mempermudah pembaca berita diperlukan pengelompokan berdasarkan topik yang dibicarakan untuk menemukan informasi yang dibutuhkan. Akan tetapi terkadang ada juga berita yang sama dituliskan dengan penambahan sedikit informasi. Hal ini menyebabkan untuk mendapatkan informasi utama dari berita tersebut menggunakan waktu yang lebih banyak. Oleh karena itu diperlukan peringkasan kumpulan berita ini agar dapat memperoleh informasi dari berita lebih mudah dan efisien.

1.2 Rumusan Rancangan

Rancangan aplikasi yang digunakan untuk clustering berita dengan menggunakan metode K-Means dan peringkasan berita dengan menggunakan metode Maximum Marginal Relevance (MMR) ini berbasis website (browser) sehingga dapat dengan mudah diakses dan digunakan. Aplikasi ini memberikan ringkasan berita yang merupakan hasil dari peringkasan beberapa

dokumen berita yang sudah dikelompokkan dan ringkasan tersebut dalam bentuk ekstraksi.

Tampilan dari aplikasi ini dirancang dengan menggunakan bahasa pemrograman ASP.Net dan sedangkan untuk perhitungan clustering berita dan peringkasan berita menggunakan bahasa pemrograman Python.

1.3 Tujuan Rancangan

Tujuan dari perancangan aplikasi clustering dan peringkasan berita ini sebagai berikut :

1. Membuat sebuah aplikasi berbasis web untuk clustering berita dengan metode K-Means.
2. Membuat sebuah aplikasi berbasis web untuk peringkasan berita dengan metode Maximum Marginal Relevance.

2. Landasan Teori

2.1 Peringkasan Teks Otomatis

Peringkasan teks Otomatis adalah proses dimana akan diambil informasi dan point-point penting dari sebuah teks sehingga menghasilkan kesimpulan dari sumber teks tersebut. Sistem akan dimasukkan berupa teks dan menghasilkan sebuah ringkasan dari teks aslinya.

2.2 *Clustering*

Clustering adalah proses untuk mengelompokkan kelas dengan kesamaan objek-objek tertentu. Clustering dan klasifikasi adalah hal yang berbeda, dimana pada clustering tidak adanya target dalam melakukan pengelompokkan. Clustering biasanya dilakukan untuk proses awal dalam data mining untuk melakukan suatu analisis. Untuk melakukan clustering, banyak algoritma yang bisa digunakan untuk clustering, salah satunya adalah K-Means. Walaupun algoritmanya berbeda-beda namun tetap mempunyai prinsip yang sama, yaitu mengelompokkan data dan mengukur kemiripan antar data dalam satu kelompok. [2]

2.3 *Text Mining*

Text mining atau penambangan teks adalah proses ekstraksi atau pemisah pola yang berupa pengetahuan dan informasi-informasi yang penting dari sumber dokumen teks seperti dokumen PDF, dokumen Word dan sejenisnya. Proses yang sering dilakukan oleh teks mining di antaranya adalah meringkas otomatis dan mendeteksi plagiarisme.

2.4 *Text Preprocessing*

Text preprocessing merupakan langkah awal dalam melakukan persiapan data teks yang tidak terstruktur menjadi lebih terstruktur untuk diolah lebih lanjut .[3] Adapun tahapan-tahapan dari proses Text Preprocessing, yaitu :

1. Segmentasi Kalimat

Pada proses segmentasi kalimat, teks berita akan dipisahkan berdasarkan tanda pemisah dalam suatu kalimat seperti tanda titik, tanda seru, dan tanda tanya yang nantinya akan digunakan untuk proses yang lebih lanjut.

2. *Case Folding*

Case Folding adalah proses dimana semua huruf dalam teks dokumen diubah menjadi huruf kecil dan semua tanda baca, nomor dan simbol akan dihilangkan.

3. *Tokenizing*

Tokenizing adalah proses dimana mengubah dari bentuk kalimat menjadi bentuk kata-kata tunggal dengan memisahkan berdasarkan spasi, hal ini bertujuan agar dapat melakukan proses *stemming.*

4. *Filtering*

Filtering yaitu proses pemilihan kata-kata penting yang bisa mewakili isi dari sebuah dokumen tersebut. Proses ini dilakukan dengan menghilangkan stopwords, yang bisa mempengaruhi hasil dari ringkasan dikarenakan bobot katanya yang besar. Stopword bisa berupa kata ganti, kata penghubung dan lain-lain. Seperti dia, yang, dari, dan lain-lain.

5. *Stemming*

Stemming merupakan proses mengubah kata menjadi kata dasar dengan menghilangkan kata-kata imbuhan seperti prefix, sufiks, infiks, dan konfiks pada setiap kata.

2.5 *Term Weighting* (Pembobotan)

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. [4]

Untuk rumus TF dapat dilihat pada Persamaan 1.

$$Tf_{t,d} = \begin{cases} 1 + \log(tf_{t,d}) & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

Dimana tf adalah jumlah dari kemunculan sebuah kata dalam satu dokumen. Sedangkan untuk rumus IDF dapat dilihat pada Persamaan 2.

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

Dimana N adalah jumlah dari semua dokumen dalam data set dan dft adalah jumlah dari dokumen yang mengandung Term t di dalamnya.

TF-IDF adalah hasil perkalian antara TF dengan IDF pada setiap katanya. Untuk rumus TF-IDF dapat dilihat pada Persamaan 3.

$$W_{t,d} = Tf_{t,d} * IDF_t = Tf_{t,d} * log\left(\frac{N}{df_t}\right) \quad (3)$$

Nilai dari TF-IDF juga dapat dilakukan normalisasi untuk rumusnya dapat dilihat pada Persamaan 4.

$$W_{t,d\ norm} = \frac{W_{t,d}}{\sqrt{\sum_{t=1}^{n} W_{t,d}{}^2}} \quad (4)$$

Dimana W adalah bobot Term t pada dokumen d, t adalah kata ke-t dari Term, d adalah dokumen ke-d, N adalah total dokumen, tf adalah jumlah kemunculan Term t pada dokumen d, df adalah jumlah dokumen yang memiliki Term t.

2.6 *Cosine Similarity*

Cosine similarity adalah tahapan dimana untuk menghitung kemiripan dari antar dokumen satu dengan dokumen lainnya. Nilai dari cosine similarity digunakan untuk mewakili kesamaan antara dokumen dalam melakukan peringkasan informasi, untuk rumus cosine similarity dapat dilihat pada Persamaan 5 dan Persamaan 6. [4]

1. Untuk Cosine Similarity tanpa normalisasi TF-IDF:

$$CosSim(d_j,q) = \frac{\sum_{i=1}^{t}(W_{i,j}.W_{i,q})}{\sqrt{\sum_{i=1}^{t} W_{i,j}{}^2 \cdot \sum_{i=1}^{t} W_{i,q}{}^2}} \quad (5)$$

2. Untuk Cosine Similarity dengan normalisasi TF-IDF:

$$CosSim(d_j,q) = \sum_{i=1}^{t}(W_{i,j}norm \times W_{i,q}norm) \quad (6)$$

Dimana i adalah Term dalam Kalimat, Wi,j adalah bobot Term i dalam blok j, Wi,q adalah bobot Term i dalam blok q.

2.7 Algoritma *Clustering K-Means*

Algoritma K-Means cukup popular dikarenakan cukup mudah untuk diimplementasikan. Adapun kelemahan dari algoritma K-Means yaitu sensitif terhadap inisialisasi cluster. Untuk algoritma dari K-Means Clustering adalah sebagai berikut: [5]

1. Inisialisasi cluster k. Biasanya dilakukan dengan cara random. Namun pada penelitian ini untuk menginisialisasi jumlah k dapat menggunakan rumus pada Persamaan 7. Dimana n adalah jumlah total dokumen. [6]

$$k = \sqrt{n/2} \quad (7)$$

2. Masukkan dokumen ke cluster berdasarkan ukuran kedekatan dengan centroid. Dimana centroid merupakan vektor yang menjadi pusat dari cluster. Untuk rumus kedekatannya dapat dilihat pada persamaan 8.

$$d(P,Q) = \sqrt{\sum_{j=1}^{p}(x_j(P) - x_j(Q))^2} \quad (8)$$

3. Setelah seluruh dokumen masuk ke cluster, maka hitung ulang centroidnya berdasarkan dokumen yang ada di dalam cluster tersebut dengan mencari nilai rata-rata centroid dari tiap cluster.

4. Jika nilai dari centroidnya berubah maka kembali ke langkah 2. Jika nilai dari centroidnya tidak berubah, maka berhenti.

2.8 *Maximum Marginal Relevance* (MMR)

Maximum Marginal Relevance atau MMR adalah salah satu metode yang digunakan untuk meringkas dokumen tunggal maupun banyak dokumen. *Maximum Marginal Relevance* merupakan Teknik peringkasan yang tidak mengandung redundansi dan mengambil informasi yang relevan. *Maximum Marginal Relevance* meringkas dokumen dengan memperhitungkan kesamaan antara query dengan kalimat isi dokumen untuk mendapatkan nilai yang nantinya akan dijadikan ringkasan. Untuk rumus dari *Maximum Marginal Relevance* dapat dilihat pada Persamaan 9.[4]

$$MMR = argmax[\lambda * Sim1(Si,Q) - (1-\lambda) * max\ Sim2(Si,S')] \quad (9)$$

Dimana λ adalah parameter yang mempengaruhi tingkat relevansi, Si adalah vektor bobot kata yang menjadi kandidat, S' adalah vektor bobot kata lainnya selain kandidat, Q adalah vektor bobot kata dari query, Sim1(Si,Q) adalah nilai similarity antar kalimat ke-i dengan query dan Sim2(Si,S') adalah nilai similarity antar kalimat ke-i dengan kalimat hasil ekstraksi.

Nilai dari parameter λ merupakan 1 atau 0 atau antara ($0 < \lambda < 1$). Jika pada saat parameter λ sama dengan 1 maka nilai *Maximum Marginal Relevance* yang diperoleh akan cenderung relevan terhadap dokumen aslinya. Sedangkan ketika λ sama dengan 0 maka nilai dari *Maximum Marginal Relevance* yang diperoleh akan cenderung relevan dengan kalimat yang sudah diekstrak sebelumnya. Oleh sebab itu, nilai dari parameter λ perlu dioptimalkan agar mendapatkan hasil ringkasan yang baik dengan nilai λ berada antara 0 sampai 1. Pada peringkasan dengan dokumen yang sedikit, seperti artikel akan menghasilkan hasil ringkasan dengan baik jika nilai dari parameter λ adalah 0,7.[4]

2.9 *Silhouette Coefficient*

Silhouette Coefficient adalah salah satu metode yang digunakan untuk menguji kualitas dari sebuah cluster. Metode ini merupakan kombinasi dari metode *Cohesion* dan metode *Separation*. Metode Cohesion adalah metode yang digunakan untuk mengukur kedekatan relasi antar objek dalam satu cluster yang sama, sedangkan metode Separation adalah metode yang digunakan untuk mengukur kejauhan sebuah cluster terpisah dengan cluster yang lainnya. *Silhouette Coefficient* mempunyai tiga tahap dalam perhitungannya, Berikut adalah tahapan untuk perhitungan Silhouette Coefficient.[3]

1. Menghitung nilai rata-rata dari jarak objek dengan semua dokumen yang ada di dalam satu cluster dengan menggunakan rumus pada persamaan 10.

$$a(i) = \frac{1}{|A|-1} \sum j \in_{A,j\neq i} d(i,j) \quad (10)$$

Dimana a(i) merupakan perbedaan rata-rata objek (i) dengan semua objek lain pada A, d(i,j) adalah jarak antara data i dengan j, dan A adalah cluster.

2. Lalu menghitung jarak dari objek dengan semua dokumen antar cluster dengan menggunakan rumus pada persamaan 11.

$$d(i,C) = \frac{1}{|C|} \sum j \ \in C \ d(i,j) \quad (11)$$

Dimana d(i,C) adalah perbedaan rata-rata objek (i) dengan semua objek lain pada C (cluster lain selain cluster A atau cluster C berbeda dengan A). Setelah menghitung d(i, C) untuk semua C, cari nilai minimum dengan menggunakan rumus pada persamaan 12.

$$b(i) = min_{C\neq A} d(i,C) \quad (12)$$

Dimana cluster B yang mencapai minimum (yaitu, d(i,B) = b(i)) disebut tetangga dari objek (i). Ini merupakan cluster terbaik kedua untuk objek (i).

3. Kemudian hitung nilai *Silhouette Coefficient* dengan menggunakan rumus pada persamaan 13.

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (13)$$

Nilai hasil *Silhouette Coefficient* mempunyai bervariasi antara -1 hingga 1. Jika nilai dari *Silhouette Coefficient* mendekati nilai 1, maka semakin baik pula pengelompokan data dalam satu cluster. Sebaliknya jika nilai dari *Silhouette Coefficient* nilai -1, maka semakin buruk pula pengelompokan data dalam satu cluster. Berikut adalah tabel tingkat kualitas struktur dari *Silhouette Coefficient.* [3]

Tabel 1 Nilai *Silhouette Coefficient*

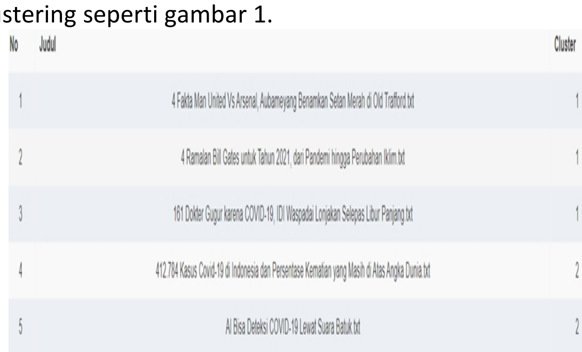
Nilai *Silhouette Coefficient*	**Struktur**
0.7 < *Silhouette Coefficient* <= 1	Kuat
0.5 < *Silhouette Coefficient* <= 0.7	Sedang
0.25 < *Silhouette Coefficient* <= 0.5	Lemah
Silhouette Coefficient <= 0.25	Tidak Terstruktur

3. Hasil Pengujian

3.1 Pengujian *Clustering*

Hasil Pengujian akurasi dari clustering menggunakan Metode Silhouette Coefficient yang nantinya akan menampilkan seberapa baik kualitas dari hasil clustering tersebut.

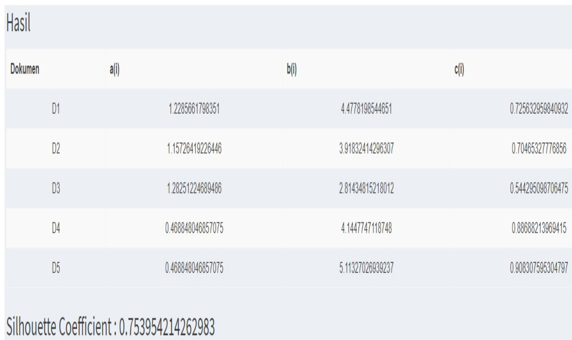
Dari berita 1 sampai 5 tersebut didapatkan hasil clustering seperti gambar 1.



No	Judul	Cluster
1	4 Fakta Man United Vs Arsenal, Aubameyang Benamkan Setan Merah di Old Trafford.txt	1
2	4 Ramalan Bill Gates untuk Tahun 2021, dari Pandemi hingga Perubahan Iklim.txt	1
3	161 Dokter Gugur karena COVID-19, IDI Waspadai Lonjakan Selepas Libur Panjang.txt	1
4	412.784 Kasus Covid-19 di Indonesia dan Persentase Kematian yang Masih di Atas Angka Dunia.txt	2
5	AI Bisa Deteksi COVID-19 Lewat Suara Batuk.txt	2

Gambar 1 Hasil Clustering Berita 1-5

Disini dijelaskan bahwa berita 1,2, dan 3 termasuk ke cluster 1. Dan berita 4 dan 5 termasuk ke cluster 2. Untuk hasil kualitas dari clusteringnya bisa dilihat pada gambar 2.



Hasil

Dokumen	a(i)	b(i)	c(i)
D1	1.2285661798351	4.4778198544651	0.725632959840932
D2	1.15726419226446	3.91832414296307	0.70465327776856
D3	1.28251224689486	2.81434815218012	0.544295098706475
D4	0.468848046857075	4.1447747118748	0.88688213969415
D5	0.468848046857075	5.11327026939237	0.908307595304797

Silhouette Coefficient : 0.753954214262983

Gambar 2 Hasil Silhouette Coefficient Berita 1-5

Pada gambar diatas, didapatkan hasil Silhouette Coefficient (SC) adalah 0.75. Maka kualitas dari clustering berita 1 sampai 5 diatas adalah kuat, karena nilai dari Silhouette Coefficient lebih besar dari 0.7. Dari hasil clustering 5 berita diatas, jika dinilai secara manual terdapat kekurangan pada saat pembagian

cluster, dikarenakan berita 2 dan berita 3 jika dilihat dari judul seharusnya masuk ke cluster ke 2.

Sedangkan untuk berita 6 sampai 10 didapatkan hasil clustering seperti pada Gambar 3.

No	Judul	Cluster
1	Anggota DPRD Sumbar Meninggal Dunia Setelah Positif Corona.txt	1
2	Arsenal Bungkam Manchester United, Taktik Arteta Disebut Mirip Mourinho.txt	1
3	Bos WHO Tidak Dites COVID-19 Usai Kontak Dekat, Ini Alasannya.txt	1
4	Cara Benar Pakai dan Lepas Sarung Tangan di Masa Pandemi Covid-19.txt	1
5	Cegah Covid-19, Hindari Menyentuh Barang-barang Ini di Mal.txt	2

Gambar 3 Hasil Clustering Berita 6 - 10

Disini dijelaskan bahwa berita 6 sampai 9 (1-4) termasuk ke cluster 1. Dan sisanya termasuk ke cluster 2. Untuk hasil kualitas dari clusteringnya dapat dilihat pada gambar 4.

Hasil

Dokumen	a(i)	b(i)	c(i)
D1	1.41313857910769	1.44750174307986	0.0237396356422069
D2	0.792136466640944	1.5326294087594	0.483151985657024
D3	0.832094117999903	1.33078870571257	0.374736113683535
D4	1.24204106380636	1.76552786445197	0.296504411618624
D5	0	3.05488157220897	1

Silhouette Coefficient : 0.435626429320278

Gambar 4 Hasil Silhouette Coefficient Berita 6 – 10

Pada gambar diatas, didapatkan hasil Silhouette Coefficient (SC) adalah 0.43. Maka kualitas dari clustering berita 6 sampai 10 di atas adalah lemah, karena nilai dari Silhouette Coefficient terletak diantara 0.25 dan 0.5. Dari hasil clustering 5 berita diatas, jika dinilai secara manual terdapat kekurangan pada saat pembagian cluster, dikarenakan pada centroid awal terletak pada berita pertama dan berita terakhir, sehingga untuk berita 2 masuk ke cluster 1.

3.2 Pengujian Peringkasan

Hasil pengujian akurasi dari hasil peringkasan menggunakan Question and Answering dikarenakan pada peringkasan tidak terdapat definisi ringkasan ideal. Maka dari itu diperlukan beberapa responden untuk menilai hasil ringkasan berita tersebut. Untuk nilai MMR dari berita 1 sampai 5 adalah sebagai berikut.

MMR

Iterasi	D1	D2	D3	D4	D5	D6	D7	D8
1	5.8410687066861	14.5389043543955	12.7058633904371	4.94459341110869	0	0	7.22459645107746	
2	0.973594132400415	7.11945217719777	3.63024668298202	1.23483948579358	-3.3396861218058	0	3.61229822553873	
3	0.64906275493361	4.74630145146518	2.42016445532135	0.823226323862386	-2.2264574145372	0	2.40819881702582	
4	0.486797066200207	3.55972608859888	1.81512334149101	0.617419742896789	-1.6698430609029	0	1.80614911276936	
5	0.389437652960166	2.84778087087911	1.45209867319281	0.493935794317432	-1.33587444872232	0	1.44491929021549	
6	0.324531377466805	2.37315072573259	1.21008222766067	0.411613161931193	-1.1132287072686	0	1.20409940851291	
7	0.278169752114404	2.03412919348508	1.03721333799486	0.352811281655308	-0.954196034801657	0	1.03208520729678	
8	0.243398533100104	1.77986304429944	0.907561670745505	0.308709871448395	-0.83492153045145	0	0.903074556384682	
9	0.216354251644537	1.58210048382173	0.806721485107115	0.274408774620795	-0.7421524715124	0	0.802732939008607	
10	0.194718826480083	1.42389043543955	0.726049336596404	0.246967897158716	-0.66793722436116	0	0.722459645107746	
11	0.177017114981894	1.29444585039959	0.660044851451276	0.224516270144287	-0.607215658510146	0	0.656781495552496	
12	0.162265688733402	1.18657536286629	0.605041113830337	0.205806580965596	-0.5566143536343	0	0.602049704256455	
13	0.149783712676987	1.0953003349535	0.558499489689541	0.189975305506704	-0.5137978648932	0	0.55573818854442	
14	0.139084876057202	1.01706459674254	0.518606668997431	0.176405640827654	-0.477098017400829	0	0.51604260364839	

Gambar 5 MMR Berita 1-5

Sedangkan untuk nilai MMR dari Berita 6-10 yaitu sebagai berikut.

MMR

Iterasi	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1	0	0	9.15778397018422	58.7296268945919	0	22.2917351580519	0	0	0	0
2	0	0	-0.926621262043721	29.214813447296	0	11.1458675790259	0	-4.03481097511725	-2.50183009096338	-3.0919347089795
3	0	0	-0.617747508029147	19.4765422981973	0	7.43057838601729	0	-2.6898739834115	-1.66788672730892	-2.06128980598633
4	0	0	-0.46331063102186	14.607406723648	0	5.57293378951297	0	-2.01740548755863	-1.25091504548169	-1.54596735448975
5	0	0	-0.370648504817488	11.6859253789184	0	4.45834703161037	0	-1.6139243900469	-1.00073203638535	-1.2367738835918
6	0	0	-0.308873754014574	9.73827114909866	0	3.71528919300864	0	-1.34493699170575	-0.83394336365446	-1.03064490299317
7	0	0	-0.264748932012492	8.34708955637028	0	3.18453359400741	0	-1.15280313574779	-0.714808597418109	-0.883409916851284
8	0	0	-0.23165531551093	7.30370336182399	0	2.78646689475648	0	-1.00870274377931	-0.625457522740845	-0.772983677244874
9	0	0	-0.205915836009716	6.49218076606577	0	2.47685946200576	0	-0.896624661137167	-0.555962242436307	-0.687096601995444
10	0	0	-0.185324252408744	5.84296268945919	0	2.22917351580519	0	-0.80696219502345	-0.500366018192676	-0.618386941795899
11	0	0	-0.168476593098858	5.31178426314472	0	2.02652137800472	0	-0.733601995475864	-0.454878198356978	-0.562169947087181
12	0	0	-0.154436877007287	4.86913557454933	0	1.85764459650432	0	-0.672468495852875	-0.41697168182723	-0.515322451496583
13	0	0	-0.142557117237496	4.49458668419938	0	1.71474885831168	0	-0.620740150018039	-0.384896937071289	-0.475682262919922
14	0	0	-0.132374466006246	4.17354477818514	0	1.5922667970037	0	-0.576401567873893	-0.357404298709054	-0.441704958425642
15	0	0	-0.123549501605829	3.89530845963946	0	1.48611567720346	0	-0.5379747966823	-0.333577345461784	-0.412257961197266
16	0	0	-0.115827657755465	3.651851680912	0	1.39323344737824	0	-0.504351371889656	-0.312728761370423	-0.386491838622437

Gambar 6 MMR Berita 6-10

Dimana dari nilai MMR maksimum tersebut akan dijadikan ringkasan. Untuk hasil ringkasan dari berita 1 sampai 5. Didapatkan 2 buah ringkasan. Dikarenakan terdapat 2 cluster berita yaitu berita 1 sampai 3 termasuk cluster 1 dan berita 4 sampai 5 termasuk cluster 2. Berikut ini adalah hasil ringkasan berita per cluster.

Tabel 2 Hasil Ringkasan Berita 1 - 5

Hasil Ringkasan Berita 1 sampai 5	
Cluster	**Ringkasan**
Cluster 1	KOMPAS Laga Man United vs Arsenal pada pekan ketujuh Liga Inggris 2020-2021 yang digelar di Stadion Old Trafford, Minggu (1/11/2020) berakhir dengan kemenangan untuk tim tamu Arsenal sukses menaklukkan Manchester United dengan skor tipis 1-0 Selain menjadi kekalahan ketiga Setan Merah di Liga Inggris musim ini, laga Manchester United vs Arsenal juga menyajikan sejumlah fakta menarik lainnya Berikut empat fakta laga Man United vs Arsenal yang dihimpun dari Opta
Cluster 2	JAKARTA, KOMPAS696 kasus baru Covid-19 dalam 24 jam terakhir Penambahan itu menyebabkan total kasus Covid-19 di Indonesia kini berjumlah 412899414 orang yang sudah diperiksa Dengan demikian, total pasien Covid-19 yang sudah sembuh dari virus corona terhitung sejak awal pandemi berjumlah 341

Sedangkan untuk hasil ringkasan dari berita 6 sampai 10 juga didapatkan 2 buah ringkasan. Karena juga terdapat 2 cluster berita yaitu berita 1 sampai 4 termasuk cluster 1 dan berita 5 termasuk cluster 2. Berikut ini adalah hasil ringkasan berita per cluster.

Tabel 3 Hasil Ringkasan Berita 6 - 10

Hasil Ringkasan Berita 6 sampai 10	
Cluster	**Ringkasan**
Cluster 1	PADANG, KOMPAS Politisi Partai Amanat Nasional (PAN) itu meninggal dunia, Sabtu (31/10/2020) pukul 23 Supardi menyebut, SF merupakan salah seorang anggota DPRD Sumbar yang aktif dan vokal dalam menyuarakan aspirasi konstituennya SF sebelum menjabat sebagai anggota DPRD Sumbar, terlebih dahulu menjadi anggota DPRD Dharmasraya Neville menilai kekuatan lini tengah menjadi kunci kemenangan Arsenal atas Manchester United "Jangan lupa, mencuci tangan dilakukan dengan air mengalir dan sabun selama 20 detik serta dikeringkan agar penggunaan sarung tangan tidak meningkatkan kelembaban dan menyebabkan jamur," kata Kesh "Saat melepaskan sarung tangan, jangan sampai merobeknya
Cluster 2	TEMPO Anggota Tim Pakar Universitas Lambung Mangkurat (ULM) untuk Percepatan Penanganan COVID-19, Prof Dr dr Syamsul Arifin MPd, mengingatkan masyarakat untuk mewaspadai transmisi COVID-19 di mal yang kerap diabaikan pengunjung "Hindari bersentuhan dengan beberapa tempat yang perlu diwaspadai jadi transmisi COVID-19 dan keharusan membawa sendiri penyanitasi tangan selama kunjungan ke mal," katanya Kemudian, pegangan kereta barang atau keranjang merupakan area yang sering disentuh oleh para pembeli pada saat membawa barang belanjaan Oleh karena itu, pada saat menyentuhnya bisa menggunakan tisu dan membersihkan tangan dengan hand sanitizer setelah selesai digunakan Terakhir untuk tombol ATM, peneliti di Cina menemukan masing-masing tombol mengandung rata-rata 1

Dari data Survey yang disebarkan menggunakan Question and Answering yang berupa kuesioner. Didapatkan total responden sebanyak 52 orang. Responden diminta untuk menjawab 5 pertanyaan yang dari tiap ringkasan sehingga total yang harus dijawab sebanyak 20 pertanyaan.

Tabel 4 Akurasi MMR

Ya Ringkasan 1	**Ya Ringkasan 2**	**Ya Ringkasan 3**	**Ya Ringkasan 4**
156	192	196	205
Avg Ringkasan 1	**Avg Ringkasan 2**	**Avg Ringkasan 3**	**Avg Ringkasan 4**
60	73.84615385	75.38461538	78.84615385
AVG Total MMR			
72.01923077			

Dari data tabel 4 dapat dilihat total yang menjawab ya pada ringkasan pertama sebanyak 156 dan rata-rata 60%, pada ringkasan kedua yang menjawab ya sebanyak 192 dan rata-rata 73.84%, pada ringkasan ketiga yang menjawab ya sebanyak 196 dan rata-rata 75.38%, dan pada ringkasan keempat yang menjawab ya sebanyak 205 dan rata-rata 78.84%. Dari keempat ringkasan tersebut dapat dilihat bahwa nilai akurasi dari MMR adalah 72.01%.

4. Kesimpulan

Berdasarkan pengujian program "Aplikasi Clustering Berita Dengan Metode K-Means Dan Peringkas Berita Dengan Metode Maximum Marginal Relevance", maka didapatkan kesimpulan sebagai berikut :

1. Pada pengujian Clustering berita 1 sampai 5 mendapatkan nilai Silhouette Coefficient yaitu 0.75, Maka kualitas dari clustering berita 1 sampai 5 diatas adalah kuat atau baik, karena nilai dari Silhouette Coefficient lebih besar dari 0.7.
2. Pada pengujian Clustering berita 6 sampai 10 mendapatkan nilai Silhouette Coefficient yaitu 0.43. Maka kualitas dari clustering berita 6 sampai 10 di atas adalah lemah, karena nilai dari Silhouette Coefficient terletak diantara 0.25 dan 0.5.
3. Pada pengujian peringkasan dari keempat ringkasan tersebut dapat dilihat bahwa nilai akurasi dari MMR adalah 72.01%.
4. Pada saat ada lebih dari 2 topik berita dari 3 sampai 5 berita, pembagian cluster tetap menjadi 2 kelompok atau cluster.
5. Batasan dari peringkasan ini ditentukan dari perolehan nilai MMR ketika nilai dari MMR maksimum dari tiap kalimat tersebut lebih besar dari 0 maka akan dijadikan ringkasan.
6. Banyaknya file dan besarnya ukuran file akan mempengaruhi lamanya proses komputasi.

REFERENSI

[1] Rofiqi, Ach. Yasir. "CLUSTERING BERITA OLAHRAGA BERBAHASA INDONESIA MENGGUNAKAN METODE K-MEDOID BERSYARAT". Jurnal Simantec. Vol. VI, Nomor 1. Jawa Timur: Universitas Trunojoyo Madura, Juni 2017.

[2] Sukma Sindi; R. O. N. Weni; Sihombing, Irma Agustika; P.P.P.A.N.W. Fikrul Ilmi R.H.Zer; dan Hartama, Dedy. "ANALISIS ALGORITMA K-MEDOIDS CLUSTERING DALAM PENGELOMPOKAN PENYEBARAN COVID-19 DI INDONESIA". Jurnal Teknologi Informasi. Vol. IV, Nomor 1. Pematangsiantar: STIKOM Tunas Bangsa Pematangsiantar, 2020.

[3] Hudin, Muhammad Sholeh; Fauzi ,M Ali; dan Adinugroho, Sigit. "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya)". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol. II, Nomor 11. November 2018.
[4] Saraswati, Nirmala Fa'izah; Indriati; dan Perdana, Rizal Setya; "Peringkasan Teks Otomatis Menggunakan Metode *Maximum Marginal Relevance* Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol. II, Nomor 11. Jawa Timur: Fakultas Ilmu Komputer Universitas Brawijaya. November 2018.
[5] Husni; D. P. N. Yudha; dan M. Syarief. "Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means". Jurnal SimanteC. Vol. IV, Nomor 3. Madura: Fakultas Teknik Universitas Trunojoyo, Juni 2015.
[6] Kodinariya, Trupti M. and Dr. Prashant R. Makwana. "Review on deTermining number of Cluster in K-Means Clustering". International Journal of Advance Research in Computer Science and Management Studies. Vol. I, Nomor 6. November 2013.

Edy Susanto, Seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara.

Viny Christanti, Seorang dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Manatap Dolok Lauro, Seorang dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.