

EKSTRAKSI FITUR SITUS BERITA ONLINE UNTUK KALEIDOSKOP BERITA TAHUNAN

Afri Yosela Putri¹, Faisal Rahutomo², Ridwan Rismanto³

^{1,2,3} Jurusan Teknologi Informasi, Program Studi Teknik Informatika, Politeknik Negeri Malang
Email: ¹ afriyoselaputri@gmail.com, ² faisal.polinema@gmail.com, ³ ridwan@polinema.ac.id

Abstrak

Informasi menjadi suatu hal yang dibutuhkan seiring dengan perkembangan teknologi informasi dan komunikasi. Salah satu sumber informasi tersebut adalah situs berita daring yang berisi artikel berita dengan topik yang berbeda. Dengan banyaknya jumlah artikel berita dengan berbagai macam topik maka proses pengelompokan tersebut menjadi sulit dilakukan dan membutuhkan waktu yang lama. Oleh karena itu, dibutuhkan sistem yang dapat mengelompokkan artikel berita secara otomatis agar proses pengelompokkan lebih mudah dan cepat. Ekstraksi fitur situs berita online bertujuan mengelempokkan artikel berita secara otomatis dan mendapatkan artikel yang populer dalam jangka waktu tertentu. Penelitian ini menggunakan tahapan *text preprocessing* untuk pengolahan teks dilanjutkan dengan pembobotan *TF IDF* pada masing-masing artikel berita sehingga terbentuk *vector term*. Berdasarkan pembobotan tersebut, artikel-artikel berita tersebut dikelompokkan dengan metode *K-Means Clustering*. Hasil pengelompokkan (*clustering*) tersebut menunjukkan jumlah populasi artikel setiap cluster. Proses pemilihan judul artikel berdasarkan kedekatan *euclidean distance*. Aplikasi ini telah diuji dengan membandingkan keluaran sistem dengan hasil keputusan manual. Pengujian tersebut dilakukan dengan masukan persentase jumlah *cluster* yang berbeda. Berdasarkan hasil pengujian tersebut, terdapat beberapa faktor yang mempengaruhi akurasi kebenaran pada pengujian metode *K-Means Clustering* yaitu persentase jumlah *cluster*, semakin besar jumlah *cluster* maka artikel berita yang dikelompokkan semakin spesifik. Hal tersebut menyebabkan tingkat akurasi kebenaran semakin tinggi.

Kata kunci: *text preprocessing*, pembobotan *TF IDF*, *vector term*, *K-Means Clustering*, *euclidean distance*.

1. Pendahuluan

Informasi menjadi suatu hal yang dibutuhkan seiring dengan perkembangan teknologi informasi dan komunikasi. Salah satu sumber informasi tersebut adalah situs berita online. Pada situs berita online, kategori berita biasanya dipisah menjadi halaman olahraga, bisnis, teknologi dan sebagainya. Semakin besar arus dokumen berita yang masuk, maka semakin luas pula sebaran topik dan kategori berita yang ada. Misalnya pada kategori "Olahraga" dibagi menjadi beberapa topik yang lebih spesifik seperti "balap motor" dan "bulu tangkis". Hal ini dapat dimanfaatkan untuk mengetahui kepopuleran topik berita pada jangka waktu tertentu atau lebih sering disebut dengan kaleidoskop berita. Untuk mengetahuinya maka diperlukan pengelompokkan artikel-artikel berita. Dengan banyaknya jumlah artikel berita dengan berbagai macam topik maka proses pengelompokan tersebut menjadi sulit dilakukan dan membutuhkan waktu yang lama karena harus melihat, membaca dan memahami isi setiap artikel berita.

Beberapa penelitian sebelumnya pernah dilakukan diantaranya Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma *K-Means* dengan hasil akurasi yang masih belum sempurna dan nilai *k (cluster)* yang masih sedikit untuk menentukan akurasi (Husni, 2015),

Pengklasifikasian Karakteristik Dengan Metode *K-Means Cluster Analysis* dengan hasil algoritma *K-Means* dapat meringkas objek dari jumlah besar sehingga lebih memudahkan untuk mendiskripsikan karakteristik dari masing-masing kelompok (Ediyanto, 2013), Perbandingan Metode Clustering Menggunakan Metode *Single Linkage* dan *K-Means* Pada Pengelompokan Dokumen dengan hasil Metode *single linkage* memiliki performansi yang lebih baik dibandingkan dengan metode *K-means* karena penelitian tersebut tidak melakukan inialisasi *cluster* awal (Rendy, 2014).

Oleh karena itu, dibutuhkan sistem yang dapat mengelompokkan artikel berita secara otomatis agar proses pengelompokkan lebih mudah dan cepat. Penelitian ini membahas aplikasi kaleidoskop berita otomatis berbahasa Indonesia. Artikel berita yang didapatkan dari situs berita daring kemudian akan dikelompokkan sesuai dengan topik berita masing-masing. Jumlah topik berita yang paling banyak akan terpilih pada kaleidoskop berita.

Sebuah kajian ilmu yang bernama *information retrieval (IR)* memunculkan beberapa metodologi yang memudahkan pencarian informasi dari sejumlah besar dokumen digital, salah satunya adalah dengan proses *clustering*. Proses tersebut adalah pengelompokan data/berkas berbasis teks berdasarkan kemiripannya. Beberapa metode *clustering* telah dikembangkan untuk

mengelompokkan data seperti *K-Means*, *decision tree*, *Naïve Bayes*, dan sebagainya. Salah satu metode *clustering*, *K-Means*, terkenal sederhana dan cepat dalam perhitungannya (Arthur, 2006), serta menjadi dasar pengembangan metode *clustering* yang lain (Kanungo, 2002; Bhatia, 2004; Pham, 2004, Mahdavi, 2008; Tarpey 2007). Metode *K-Means* yang dipadukan dengan pembobotan TF-IDF menjadi solusi untuk pengelompokan data tak terstruktur seperti dokumen teks secara otomatis. Karena TF-IDF sendiri juga merupakan metode yang populer dan memiliki hasil perhitungan yang cukup akurat (Ramos, 2010).

2. Text Preprocessing

Text preprocessing merupakan proses mempersiapkan teks dokumen atau *dataset* yang berfungsi untuk mengubah data teks yang tidak terstruktur menjadi data terstruktur. Secara umum proses yang dilakukan dalam tahapan *text preprocessing* diantaranya adalah *case folding*, *tokenizing*, dan *filtering*.

3. Pembobotan TF IDF

Metode *TF-IDF* (Robertson, 2004) merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, sederhana dan memiliki hasil yang akurat (Ramos, 2010). Metode ini menghitung nilai *Term Frequency (TF)* dan *Inverse Document Frequency (IDF)* pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini menghitung bobot setiap token t di dokumen d dengan rumus:

$$W_{(t,f)} = TF_{(t,d)} \times IDF_{(t)} \quad (1)$$

Dimana nilai $IDF_{(t)}$ didapat dari:

$$IDF_{(t)} = \log\left(\frac{|D|}{df_{(t)}}\right) \quad (2)$$

Keterangan:

$TF_{(t,d)}$: Jumlah kemunculan token t pada dokumen d

$IDF_{(t)}$: Nilai *IDF* token t

$df_{(t)}$: Jumlah dokumen yang memuat token t

$|D|$: Jumlah dokumen dalam korpus

4. K Means Clustering

Metode *K-Means clustering* merupakan metode *clustering* yang dikenalkan oleh (Lloyd, 1982). *K-Means* adalah metode *clustering* yang mengelompokkan semua data yang dimiliki ke dalam k *cluster*, dimana nilai k sudah ditentukan sebelumnya. *K-Means* mengelompokkan data berdasarkan jarak dari tiap dokumen ke pusat *cluster* (*centroid*) yang sudah ditentukan sebanyak k , dan mengelompokkan data-data ke pusat *cluster* yang terdekat.

Algoritma dari metode *K-Means* adalah sebagai berikut:

1. Pilih secara acak vektor dokumen yang akan digunakan sebagai *centroid* awal sebanyak k .
2. Cari *centroid* yang paling dekat dari setiap dokumen.
3. Hitung ulang untuk menentukan *centroid* baru dari setiap *cluster*.
4. Lakukan langkah 2 dan 3 hingga *centroid* tidak mengalami perubahan lagi.

Rumus perhitungan *centroid* baru dari setiap *cluster* dicari dengan menggunakan rumus:

$$M_k = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N_k} x_{ik} \quad (3)$$

Dengan:

M_k : Nilai *centroid* dari suatu *cluster*

N_k : Jumlah dokumen yang berada dalam satu *cluster*

x_{ik} : Nilai x dari sampel dokumen ke- i yang termasuk *cluster* k (C_k)

Sedangkan untuk menemukan jarak dua dokumen digunakan rumus

Euclidean distance:

$$d_{(i,j)} = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)} \quad (4)$$

Dengan:

$d_{(i,j)}$: Jarak dokumen ke- i ke dokumen ke- j

$x_{i(n)}$: Kata ke n di dokumen ke- i

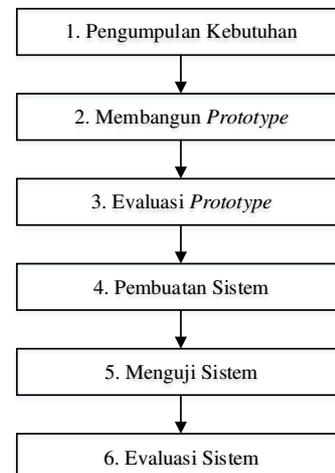
$x_{j(n)}$: Kata ke n di dokumen ke- j

Untuk mengaplikasikan *K-Means* dalam *clustering* dokumen teks, maka dibentuklah vektor dokumen dengan jumlah dimensi sebanyak token unik dalam korpus.

5. Metodologi Penelitian

Pada bagian ini dibahas metodologi yang digunakan peneliti dalam pembuatan Ekstraksi Fitur Situs Berita Online. Metode penelitian yang digunakan adalah metode *prototype*.

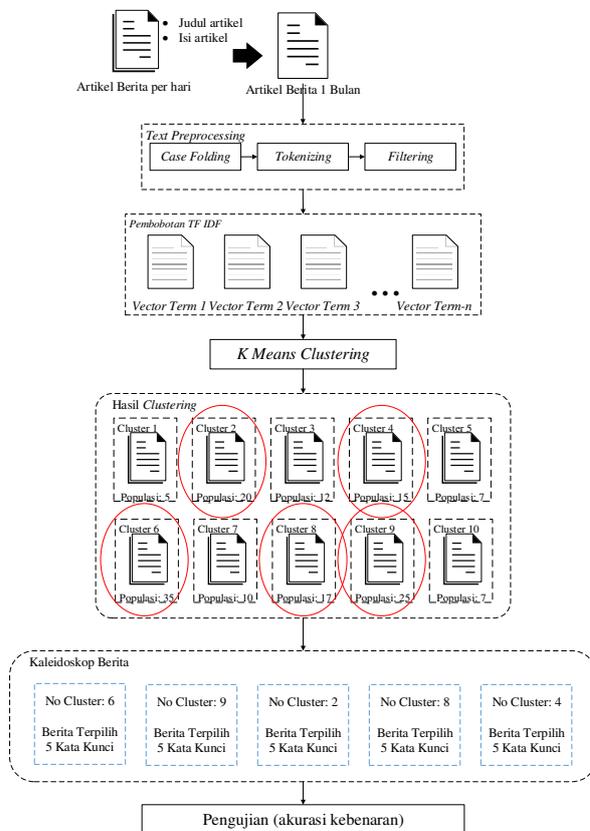
Proses kegiatan yang ada pada metode *prototyping* dapat dijelaskan pada Gambar 1.



Gambar 1. Model *Prototype*

6. Rancangan

Sistem yang dirancang dan dibangun dalam penelitian ini adalah Ekstraksi Fitur Situs Berita Online Untuk Kaleidoskop Berita Tahunan. Artikel berita yang didapatkan pada situs berita online dalam waktu enam bulan merupakan data inputan. Artikel-artikel berita tersebut disimpan dalam bentuk dokumen teks kemudian diolah dengan *text preprocessing* sehingga terbentuk token/term. Selanjutnya masing-masing token/term dihitung bobotnya dengan metode pembobotan TF IDF. Setelah itu dokumen tersebut dikelompokkan dengan *K Means Clustering*, *cluster* yang memiliki populasi terbanyak akan diambil sebagai trend topik berita atau kaleidoskop berita.

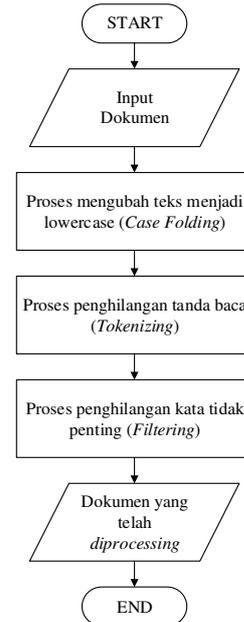


Gambar 2. Rancangan Sistem

7. Flowchart

7.1. Flowchart Text Preprocessing

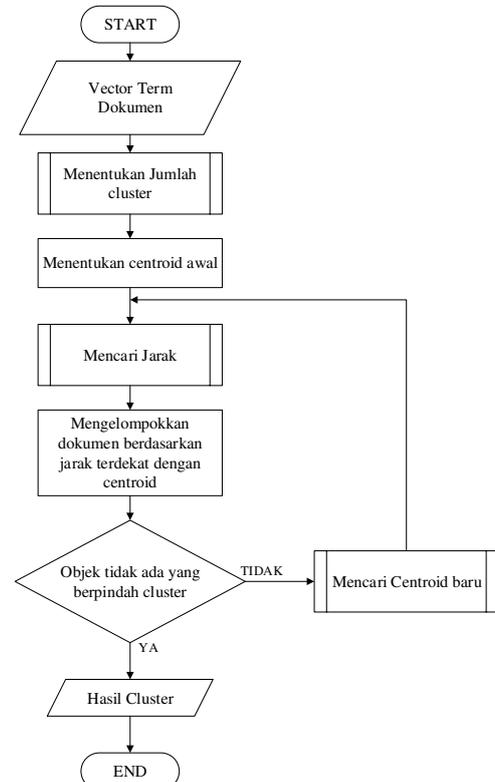
Flowchart tahap preprocessing merupakan flowchart yang berisi proses penghilangan tanda baca (*tokenization*) serta proses penghilangan kata yang tidak penting (*stopwords*).



Gambar 3. Flowchart Text Preprocessing

7.2. Flowchart K Means Clustering

Flowchart Algoritma K-Means merupakan flowchart yang berisi urutan inputan berupa vector term dari masing-masing dokumen yaitu bobot yang dimiliki setiap kata/term, mencari jumlah cluster, menentukan centroid (titik pusat) awal, mencari jarak, mengelompokkan dokumen berdasarkan jarak terdekat dengan centroid, serta proses mencari centroid baru.

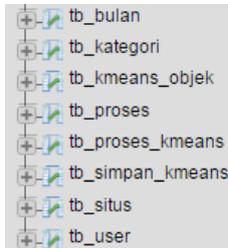


Gambar 4. Flowchart K-Means Clustering

8. Implementasi

8.1. Implementasi Basis Data

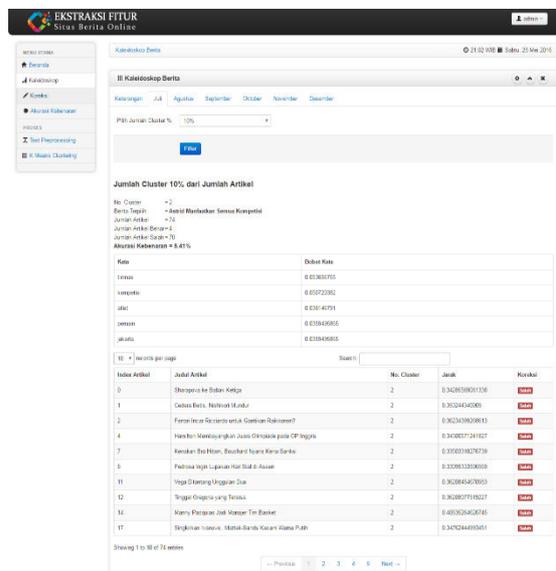
Berikut ini merupakan hasil implementasi basis data:



Gambar 5. Implementasi Basis Data

8.2. Implementasi Antarmuka

Berikut ini merupakan hasil implementasi antarmuka sistem:



Gambar 6. Implementasi Antarmuka

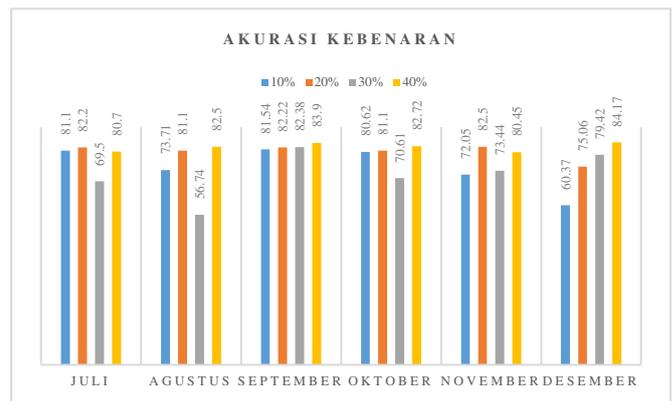
9. Uji Coba

Untuk menguji metode *K-Means Clustering* pada dokumen teks berita, maka dibutuhkan responden untuk mengoreksi kebenaran dari suatu artikel berita. Kebenaran yang dimaksud adalah kebenaran hasil clustering, apakah memiliki keterkaitan dengan artikel berita yang lain dalam satu cluster. Langkah pertama dari pengujian ini adalah melihat artikel berita terpilih, yaitu artikel berita yang mempunyai jarak *euclidean* terdekat dengan centroid. Selain itu juga terdapat lima buah kata kunci yang memiliki keterkaitan dengan artikel berita yang lain. Lima kata kunci tersebut dipilih berdasarkan bobot *TF IDF* yang paling besar.

Langkah berikutnya adalah koreksi terhadap artikel berita dalam satu *cluster*. Koreksi tersebut terdapat dua pilihan yaitu benar atau salah.

Kemudian artikel berita yang benar dijumlahkan sehingga mendapatkan jumlah artikel yang benar untuk selanjutnya dihitung menjadi persentase akurasi kebenaran artikel.

Berdasarkan pengujian manual, didapatkan perhitungan akurasi kebenaran artikel berita pada setiap persentase jumlah cluster. Semakin besar persentase jumlah *cluster* semakin besar pula tingkat akurasi kebenaran. Pada pengujian dengan persentase jumlah cluster 40% menunjukkan hasil akurasi kebenaran diatas 80%. Berikut ini adalah grafik akurasi kebenaran artikel berita pada setiap

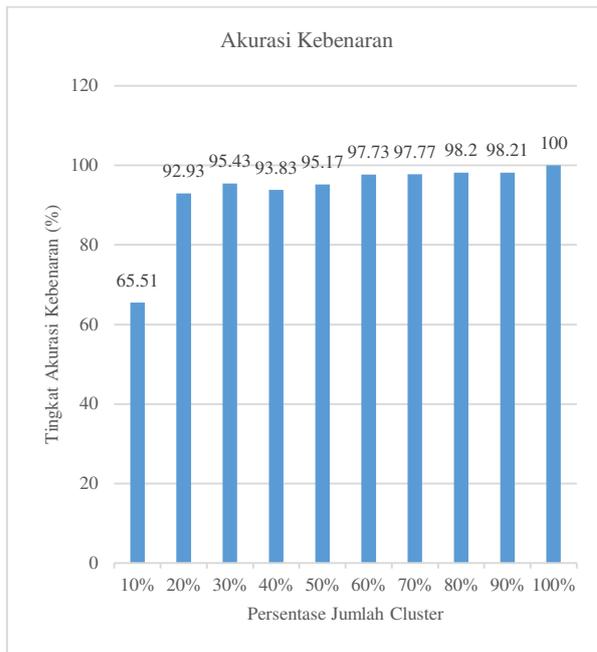


Gambar 7. Grafik Akurasi Kebenaran

bulan.

Berdasarkan grafik akurasi kebenaran penelitian, menunjukkan bahwa semakin besar persentase jumlah cluster belum tentu bisa mendapatkan tingkat akurasi yang tinggi. Hal ini disebabkan karena metode *k means clustering* mengambil titik pusat (*centroid*) secara acak. Sehingga apabila ada artikel berita yang keberagaman topiknya berjumlah sedikit menyebabkan artikel tersebut masuk ke dalam kelompok (*cluster*) yang kurang sesuai. Sebagai contoh topik tentang tenis yang berjumlah sedikit dapat masuk ke dalam topik bulu tangkis.

Untuk membuktikan hal tersebut, maka dilakukan pengujian *k means clustering* pada 47 artikel berita secara acak. Hasil dari pengujian tersebut adalah semakin besar jumlah persentase jumlah cluster, maka artikel berita akan dikelompokkan lebih spesifik. Sehingga topik yang memiliki sedikit akan dikelompokkan sendiri, hal ini akan mempengaruhi akurasi kebenaran pada masing-masing pengujian. Oleh karena itu semakin besar persentase jumlah *cluster* akan semakin tinggi akurasi kebenarannya. Berikut ini adalah grafik yang menunjukkan hasil pengujian *k-means clustering*.



Gambar 8. Grafik Akurasi Kebenaran 2

9.1. Pengujian *Text Preprocessing*

Pada tahap *text preprocessing*, artikel berita dilakukan proses *case folding*, *tokenizing*, dan *filtering*. Setelah proses tersebut dijalankan, kemudian setiap kata/term dilakukan pembobotan *TF IDF* sehingga terbentuk *vector term*. Kedua tahap tersebut dieksekusi bersama pada data berita setiap bulan. Berikut hasil eksekusi dari masing-masing data berita setiap bulan:

Bulan	Jumlah Artikel	Jumlah Term/Kata	Waktu Eksekusi
Juli	236	5562	19.104 detik
Agustus	286	6964	31.256 detik
September	195	5516	17.311 detik
Oktober	270	6837	30.635 detik
November	198	5647	18.941 detik
Desember	159	5112	14.162 detik

Tabel 1. Pengujian *Text Preprocessing*

Sumber: Pengujian *Text Preprocessing*

9.2. Pengujian *K-Means Clustering*

Pada tahap *K-Means Clustering*, *vector term* yang dihasilkan dari pembobotan *TF IDF* pada setiap artikel berita akan dikelompokkan dengan metode tersebut. Pengelompokkan (*clustering*) tersebut berdasarkan pada pendekatan *euclidean distance* dengan jumlah *cluster* sebagai data masukan. Jumlah *cluster* didapatkan dari persentase jumlah *cluster* dikalikan dengan jumlah artikel. Pada penelitian ini, terdapat empat percobaan persentase jumlah *cluster* pada setiap

clustering data berita. Berikut ini adalah hasil dari eksekusi *k-means clustering* pada masing-masing data berita setiap bulan:

Tabel 2. Pengujian *K-Means Clustering*

Bulan	Persentase Jumlah Cluster	Jumlah Cluster	Waktu Eksekusi
Juli	10%	23	286.762 detik
Juli	20%	47	545.273 detik
Juli	30%	70	411.736 detik
Juli	40%	94	555.109 detik
Agustus	10%	28	791.729 detik
Agustus	20%	57	681.101 detik
Agustus	30%	85	2202.319 detik
Agustus	40%	114	1342.303 detik
September	10%	19	188.591 detik
September	20%	39	271.223 detik
September	30%	58	406.928 detik
September	40%	78	537.713 detik
Oktober	10%	27	642.373 detik
Oktober	20%	54	753.955 detik
Oktober	30%	81	920.306 detik
Oktober	40%	108	1486.462 detik
November	10%	19	186.045 detik
November	20%	39	525.308 detik
November	30%	59	439.768 detik
November	40%	79	428.342 detik
Desember	10%	15	112.702 detik
Desember	20%	31	226.568 detik
Desember	30%	47	267.841 detik
Desember	40%	63	431.681 detik

Sumber: Pengujian *K-Means Clustering*

10. Kesimpulan dan Saran

Berdasarkan hasil dan pembahasan, maka dapat diperoleh beberapa kesimpulan diantaranya, terdapat faktor-faktor yang mempengaruhi akurasi kebenaran pada pengujian metode *K-Means Clustering* yaitu: (1) Persentase jumlah *cluster*, semakin besar jumlah *cluster* maka artikel berita yang dikelompokkan semakin spesifik. Hal tersebut menyebabkan tingkat akurasi kebenaran semakin tinggi. (2) Banyaknya keberagaman topik berita, jika jumlah keberagaman topik berita yang sedikit dikelompokkan dengan jumlah *cluster* yang kecil maka artikel berita tersebut akan masuk dalam kelompok (*cluster*) yang kurang sesuai.

Saran yang diberikan untuk pengembangan ekstraksi fitur situs berita online ini adalah dengan mencoba teknik clustering dengan pendekatan cosine similarity karena terbukti dengan menggunakan pendekatan euclidean distance hasil akurasi kebenarannya masih dipengaruhi oleh banyak faktor penentu sehingga masih belum dapat menentukan pembatasan jumlah *cluster* secara pasti untuk clustering.

Daftar Pustaka:

- Arthur, David and Sergei Vassilvitskii, 2006, How Slow is the k-Means Method, Stanford University, Stanford, CA.
- Bhatia, Sanjiv K., 2004, Adaptive K-Means clustering, Department of Mathematics & Computer Science, University of Missouri – St. Louis.
- Ediyanto. Dkk, 2013. "Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis". Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster). 2(2), 133-136
- Handoyo, Rendy. Dkk, 2014. "Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K - Means Pada Pengelompokan Dokumen". ISSN. 1412-0100. 15(2), 73-82
- Husni, Yudha Dwi Putra Negara, dkk, 2015. "Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means". Jurnal Cimantec. 4(3), 159-166
- Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.
- Mahdavi, Mehrdad and Hassan Abolhassani; 2008, Harmony k-Means Algorithm for Document clustering, Springer Science+Business Media, LLC 2008.
- Ramos, Juan, 2010, Using TF-IDF to Determine Word Relevance in Document Queries, Department of Computer Science, Rutgers University, Piscataway.
- Robertson, Stephen, 2004, Understanding Inverse Document Frequency: On Theoretical Arguments for IDF, Journal of Documentation; 2004; 60, 5; ABI/INFORM Global.
- Tarpey, Thaddeus, 2007, A Parametric k-Means Algorithm, © Springer Verlag 2007, Computational Statistic 22: 71-89.