KLASIFIKASI DOKUMEN BERITA BERBAHASA INDONESIA MENGGUNAKAN METODE *NsAIVE BAYES CLASSIFIER (NBC)* DAN *K-MEANS CLUSTERING*

Riani[1], Amir Hamzah[2], Erna Kumalasari N[3]

[1,2,3] Teknik Informatika, Institut Sains & Teknologi AKPRIND, Yogyakarta
Riani2833@gmail.com, amir@akprind.ac.id, ernakumala@akprind.ac.id

ABSTRACT

News is one of the important needs for people in various parts of the world. Through the news, the public can know the various information that is happening in the community such as economic, political, health, criminal, and natural disasters. Increasing information needs every day to make some agencies to be able to present the news quickly, accurately, reliably, and accurately through print and electronic media that can be enjoyed by news readers. Increasing the amount of news gained every day resulted in large data accumulation of text documents either online and offline. This makes it difficult in searching and classifying documents as needed.

To simplify the classification of text document news in Indonesian, one of them by using the method of NBC and K-Means Clustering. Where the calculation of NBC done not randomly, while the calculation for K-Means Clustering done randomly in this study the calculation is done 5 time on each document so the accuracy result is less accurate when compared with NBC

The highest accuracy result obtained from the research of news classification with nbc method is 50% and the average for the whole of the four documents is 45.33%, while the highest accuracy result of the classification using k_means method is 100% and the overall result is 53.26% with details Document_0 obtained an average yield of 54%, document 1 the average yield of 60%, document 2 the overall average yield of 42.56%, and document 3 the overall average yield of 52.48%.

INTISARI

Berita merupakan salah satu kebutuhan penting bagi masyarakat diberbagai belahan dunia.Melalui berita, masyarakat dapat mengetahui berbagai informasi yang sedang terjadi dimasyarakat saperti ekonomi, politik, kesehatan, kriminal, maupun bencana alam.Kebutuhan informasi yang meningkat setiap harinya membuat beberapa pihak instansi untuk dapat menyajikan berita secara cepat, tepat, terpercaya, dan akurat melalui media cetak maupun elektronik yang dapat dinikmati oleh pembaca berita. Meningkatnya jumlah berita yang didapat setiap harinya mengakibatkan penumpukan data yang besar berupa dokumen teks baik secara *online* maupuun *offline.* Sehingga menyulitkan dalam pencarian dan pengklasifikasian dokumen yang sesuai dengan kebutuhan.

Untuk mempermudah dalam pengklasifikasian dokuumen teks berita Berbahasa Indonesia salah satunya dengan menggunakan metode *NBC* dan *K-Means Clustering*. Dimana perhitungan *NBC* dilakukan tidak secara random, sedangkan perhitungan untuk *K-Means Clustering* dilakukan secara random, dalam penelitian ini perhitungan *k-means clustering* dilakukan 5x pada setiap dokumen sehingga hasil akurasinya kurang akurat jika dibandingkan dengan *NBC*

Hasil akurasi tertinggi yang diperoleh dari penelitian pengklasifikasian berita dengan metode *nbc* sebesar 50% dan rata-rata untuk keseluruhanya dari keempat dokumen sebesar 45.33%, sedangkan hasil akurasi tertinggi pengklasifikasian menggunakan metode *k_means* sebesar 100% dan hasil rata-rata keseluruhannya sebesar 53.26% dengan rincian dokumen_0 diperoleh hasil rata-rata sebesar 54%, dokumen 1 hasil rata-rata sebesar 60%, dokumen 2 hasil rata-rata keseluruhannya sebesar 42.56%, dan dokumen 3 hasil rata-rata keseluruhannya sebesar 52.48%.

PENDAHULUAN

Di era modernisasi saat ini, berita merupakan kebutuhan informasi yang sangat penting bagi masyarakat. Melalui berita, berbagai informasi yang sedang terjadi dimasyarakat dapat diketahui seperti olahraga, politik, ekonomi, kriminal, kesehatan, maupun bencana alam. Karena kebutuhan masyarakat akan berita semakin meningkat setiap harinya, berbagai pihak mencoba untuk menyajikan sajian berita berbahasa Indonesia yang sesuai dengan kebutuhan masyarakat Indonesia secara cepat, tepat, akurat, dan terpercaya bagi para pembaca berita. Berita dapat diperoleh dari media cetak maupun elektronik seperti koran, televisi, radio, dan internet. Penyampaian berita saat ini tidak hanya melalui media cetak saja melainkan sudah banyak berkembang penyampaian informasi berita melalui media internet seperti *Google News, Detik.com. dan kompas.com* sebagai sarana penyajian berita bagi para pembaca berita.

Dengan melimpahnya jumlah berita yang cukup besar setiap harinya mengakibatkan terjadinya penumpukan data berita berupa dokumen teks, baik secara *online* maupun secara *offline*. Dokumen teks berita yang menumpuk menyebabkan sulitnya untuk mencari dokumen yang sesuai dengan kebutuhan. Jika penumpukan dokumen berita masih dalam skala rendah proses pencarian dokumen berita dapat dilakukan secara manual, tapi jika penumpukan dokumen berita dalam jumlah besar proses pencarian dokumen berita tidak dapat dilakukan secara menual karena merepotkan dan menghabiskan banyak waktu yang terbuang sia - sia. Penumpukan dokumen berita juga mengakibatkan kelemahan pada proses *Text Mining*, dimana dapat menyulitkan kata yang diambil serta mengurangi ketepatan kata pada proses klasifikasi. Untuk mempermudah pada proses klasifikasi, maka dapat dibuat sebuah aplikasi menggunakan Java karena penggunaanya mudah dan tampilannya sederhana, sehingga bermanfaat untuk menyelesaikan beberapa masalah yang ada. Selain itu dibutuhkan adanya sebuah metode *Data Mining, Data Mining* merupakan suatu proses untuk menambang atau menemukan pola tertentu pada jumlah data yang besar. Teknik yang digunakan pada proses klasifikasi ini menggunakan algoritma *(NBC)* yaitu merupakan suatu teknik algoritma klasifikasi yang menggunakan metode *probabilitas,* dan membandingkannya dengan *K-Means Clustering* yaitu merupakan teknik algoritma *clustering* terhadap banyaknya data *cluster* yang diingink an dan menetapkan nilai – nilai klasifikasi secara random.

*T*INJAUAN PUSTAKA

Dalam penelitian ini digunakan beberapa referensi yang berhubungan dengan obyek penelitian. Adapun referensi itu diambil dari penelitian sebelumnya yang berhubungan dengan penelitian ini, diantaranya:

Referensi pertama, (Riani,2016 .), penelitian yang berkaitan dengan pengelompokkan opini menjadi opini negatif dan opini positif yang diambil dari opini mahasiswa terhadap sarana dan prasarana di kampus yang dikelompokkan dengan algoritma *NBC*, dimana pengujian menggunakan 4 koleksi dokumen yang masing-masing dokumen terdapat jumlah opini yang berbeda-beda. Koleksi dokumen 1 berisi 10 opini, koleksi dokumen 2 berisi 30 opini, koleksi dokumen 3 berisi 40 opini, dan koleksi dokumen 4 berisi 50 opini. Diperoleh hasil akurasi tertinggi 86.67% dan akurasi rata-rata keseluruhan dokumen yang diujikan diperoleh 83.41%.

Referensi kedua diambil dari Jurnal Ilmiah Teknik Industri, Vol.12, No.1,2013, Implementasi Algoritma K-means Clustering Untuk Menentukan Strategi Marketing President University, penelitian yang berkaitan dengan media promosi kepada masyarakat khususnya calon mahasiswa tentang kualitas dan kuantitas diperguruan tinggi yang banyak diminati oleh calon mahasiswa berdasarkan tingkat kemampuan akademik

Referensi ketiga, diambil dari Jurnal Ilmiah Semesta Teknika, Vol, 18, No.1, Hal 76-82. Penelitian yang berkaitan tentang pengelompokan mahasiswa teknik informatika UMM Magelang yang dapat mewakili kampus untuk mengikuti lomba berdasarkan ipk tertinggi dan atribut-atribut mata kuliah menggunakan Weka Interface dengan metode *K-Means Clustering.*

Berdasarkan referensi diatas maka dalam penelitian tentang Klasifikasi Dokumen Berita Berbahasa Indonesia Menggunakan Metode *NBC* dan *K-Means Clustering* perlu dilakukan untuk mempermudah pengguna dalam mengelompokkan berita yang diinginkan secara

otomatis, dalam penelitian ini berita yang akan dikelompokkan berisi berita bisnis dan berita kesehatan yang diambil melalui situs www.TribunNews.compada bulan Februari 2017 sebanyak 68 berita yang akan dikelompokkan dengan metode *NBC* dan *K-Means Clustering* sebagai tingkat perbandingan dari hasil perolehan akurasi pada setiap metode.

Landasan teori yang digunakan dalam penelitian ini adalah penjelasan teori dari penelitian sebelumnya yang berhubungan dengan penelitian ini, diantaranya:

Data Mining

Han, J., Kamber, M., & Pei, J. (2012) mengatakan bahwa *Data Mining* adalah proses menentukan pola dan informasi dari data yang berjumlah besar. Sumber data dapat berupa *database, data warehouse, web,* tempat penyimpanan informasi lainnya atau data yang mengalir kedalam sistem yang dinamis

Text Mining

Menurut Budi Susanto *teks mining* adalah penerapan konsep dan teknik data mining untuk mencari pola dalam teks.

Naive Bayes Classifier (NBC)

ilmuwan Inggris Thomas Bayes mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik

K-Means Clustering

Menurut Sarwono *algoritmak-means* adalah algoritma yang mempartisi data kedalam *cluster-cluster* sehingga data yang memiliki kemiripan berada pada satu *cluster* yang sama dan data yang memiliki ketidaksamaan berada pada *cluster* yang lain.

METODOLOGI PENELITIAN

Metode pengumpulan data yang dilakukan dalam penelitian ini adalah:

1. Observasi

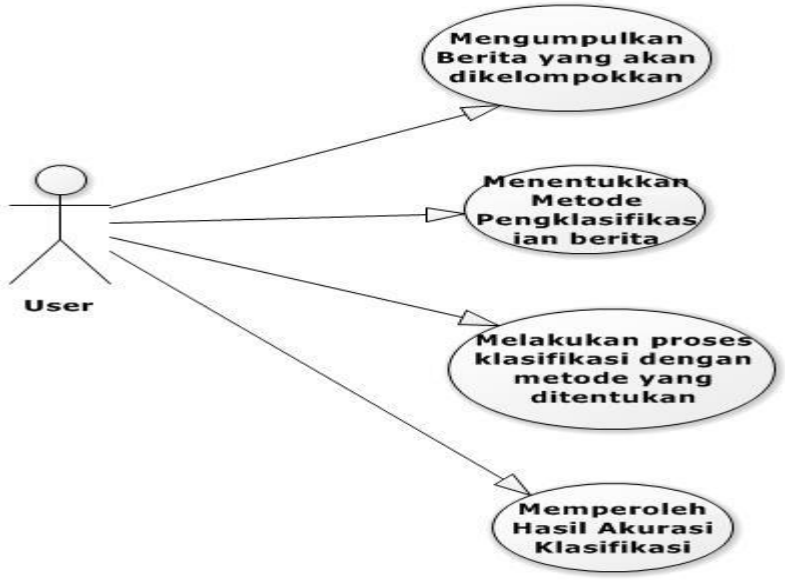
   Observasi yang dilakukan dalam penelitian ini adalah pengenalan sistem yang akan dibuat terkait dengan kelemahan dan kelebihan metode yang digunakan.

2. Pengumpulan Data Penelitian

   Pengumpulan data yang dilakukan pada penelitian ini adalah mencari data berita berbahasa Indonesia melalui internet di situs www.TribunNews.compada bulan februari 2017 sebanyak 68 berita yang terdiri dari berita bisnis dan berita kesehatan.

Alat yang digunakan dalam penelitian ini adalah perangkat keras dan perangkat lunak dengan spesifikasi sebagai berikut:

1. *Hardware* : *Processor* Intel(R) Core i3, *Memory* 2 GB, *Hardisk* 640 GB, CPU @ 2.00 GHz, Layar 14 inch LED HD.
2. *Software* : Sistem operasi yang digunakan windows 7 Pro 32 bit, Bahasa pemrograman yang digunakan adalah java dengan NetBeansIDE versi 8.0.2, Koneksi database dengan XAMPP versi 3.2.2, Perancangan *flowchart* menggunakan Software Ideas Modeler versi 10.68



Gambar 1. Use Cace Diagram

PEMBAHASAN

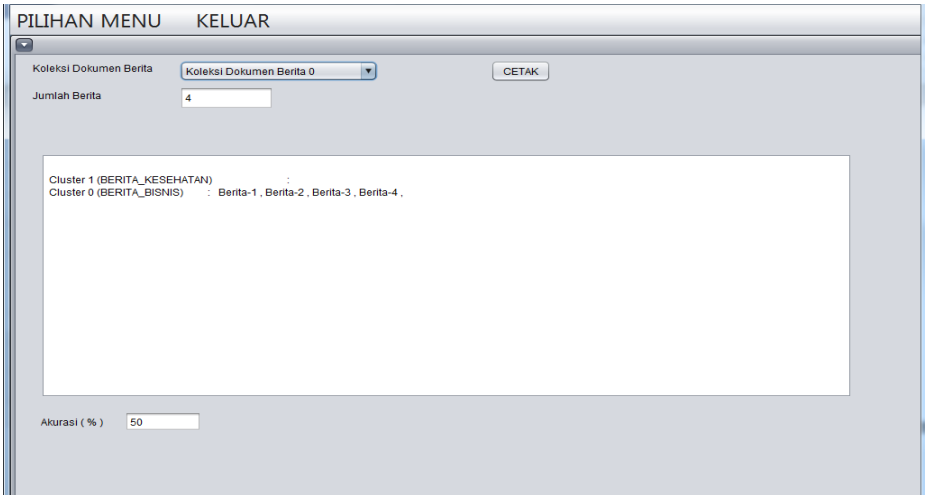
Langkah pertama yang dilakukan dalam penelitian ini untuk memperoleh hasil akurasi berdasarkan data berita yang telah dikelompokkan sebelumnya, yang terdiri dari berita bisnis dan berita kesehatan melalui situs www.TribunNews.com, lalu memilih metode yang digunakan untuk melakukan proses meng*cluster* berita

Dalam penelitian ini, metode yang digunakan adalah metode *nbc d*an metode *k-means clustering*. Contoh yang diambil adalah koleksi dokumen berita_0 yang berisi 4 berita, 2 berita bisnis dan 2 berita kesehatan yang telah ditentukan class_beritanya, *cluster* 0 untuk berita bisnis dan *cluster* 1 untuk berita kesehatan, ditunjukan pada Tabel 1.

Tabel 1. Koleksi dokumen berita_0

NO	Id_berita	Judul_berita	Isi_berita	Class_berita
1	Berita-01	Ekonomi Bisnis Terpadu	Bisnis Ok 100% Bisni untung Bisnis Rugi 5%	0
2	Berita-02	Kesehatan Rambut	Ada beberapa hal yang perlu dilakukan untuk menjag:	1
3	Berita-03	Kesehatan Kulit	Kulit lembut kulit bercahaya kulit pecah-pecah	1
4	Berita-04	Bisnis Mendunia	bisnis lancar bisnis terpadu bisnis kurang mengu	0

Berdasarkan koleksi dokumen berita_0 diatas, maka akan dilakukan proses perhitungan hasil akurasi menggunakan metode *nbc* yang dihitung secara tidak acak dan metode*k-means clustering* yang perhitungannya secara acak (random), ditunjukkan oleh Gambar 2. Hasil *cluster nbc*, Gambar 3. Hasil *cluster k-means clustering*, Gambar 4.Grafik akurasi *nbc* dan *k-means clustering,*Tabel 2. Hasil akurasi koleksi dokumen berita_0 dengan metode *nbc*, dan Tabel 3. Hasil akurasi koleksi dokumen berita_0 dengan metode *k-means clustering,* Tabel 4.Analisis akurasi klasifikasi berita



Gambar 2. Hasil cluster nbc

Keterangan Gambar 2.:

Hasil *cluster nbc* pada dokumen_0 yang berisi 4 data berita terdiri dari 2 berita kesehatan yaitu Berita-2, Berita-3, dan 2 berita bisnis yaitu Berita-1, Berita-4, diperoleh hasil akurasi sebesar 50 % artinya, ketepatan data dalam mengelompokkan berita belum sepenuhnya benar karena masih terdapat 2 berita yang salah letaknya yaitu berita kesehatan yang terdiri dari Berita-2, Berita-3, yang seharusnya masuk kedalam cluster 1 (BERITA KESEHATAN) tetapi malah masuk kedalam cluster 0 (BERITA BISNIS). Hal itu yang menyebabkan rendahnya tingkat ketepatan data yang diperoleh dari hasil jumlah data berita yang benar / jumlah dokumen berita yang diuji * 100%.

Tabel 2. Hasil akurasi koleksi dokumen_0 dengan metode *nbc*

No	Id_berita	Class berita	Hasil cluster
1	Berita-01	0	0
2	Berita-02	1	0
3	Berita-03	1	0
4	Berita-04	0	0

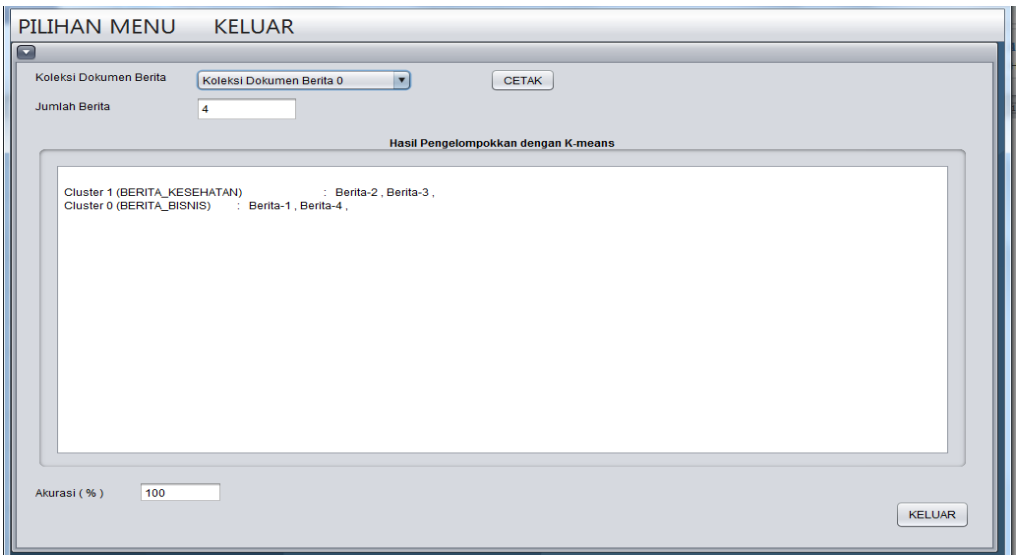
Perhitungan akurasi menggunakan *nbc* pada koleksi dokumen_0 (berisi 4 data berita)

$$\text{Akurasi} = \frac{\text{hasil data pengelompokkan}}{\text{jumlah total data berita}} \times 100\%$$

$$= \frac{2}{4} \times 100\% = 50\%$$

Keterangan Tabel 2.:

Hasil akurasi sebesar 50% artinya ketepatan data yang diperoleh belum sepenuhnya benar karena masih ada 2 berita yang letaknya salah, tidak sesuai tempatnya. Seharusnya hasil cluster berita-02, dan berita-03 masuk ke cluster 1 yaitu berita kesehatan, tetapi malah masuk ke cluster 0 yaitu berita bisnis semua, sehingga hasilnya masih rendah dibawah 75%.



Gambar 3.Hasil *cluster k-means clustering*

Keterangan Gambar3 :

Hasil*clusterk-means clustering* pada dokumen_0 yang berisi 4 data berita terdiri dari 2 berita kesehatan yaitu Berita-2, Berita-3, dan 2 berita bisnis yaitu Berita-1, Berita-4, diperoleh hasil akurasi sebesar 100 % artinya, ketepatan data dalam mengelompokkan berita sudah benar karena tingkat ketepatan data yang diperoleh sudah diatas 75% dan tidak ada letak berita yang salah.

Tabel 3. Hasil akurasi koleksi dokumen_0 dengan metode *k-means clustering*

No	Id_berita	Class berita	Hasil cluster
1	Berita-01	0	0
2	Berita-02	1	1
3	Berita-03	1	1
4	Berita-04	0	0

Perhitungan akurasi menggunakan *nbc* pada koleksi dokumen_0 (berisi 4 data berita)

Akurasi = $\frac{\text{hasil data pengelompokkan}}{\text{jumlah total data berita}} \times 100\%$

$= \frac{4}{4} \times 100\% = 100\%$

Keterangan Tabel 3.:

Hasil akurasi tertinggi sebesar 100% karena metode *k-means clustering* perhitungan dilakukan secara random dan hasilnyya bisa berubah.artinya ketepatan data yang diperoleh sudah tepat karena data berita sudah sesuai letaknya, sehingga hasil akurasinya sudah sempurna dan diatas 75%.

Tabel 3. Analisis akurasi klasifikasi berita

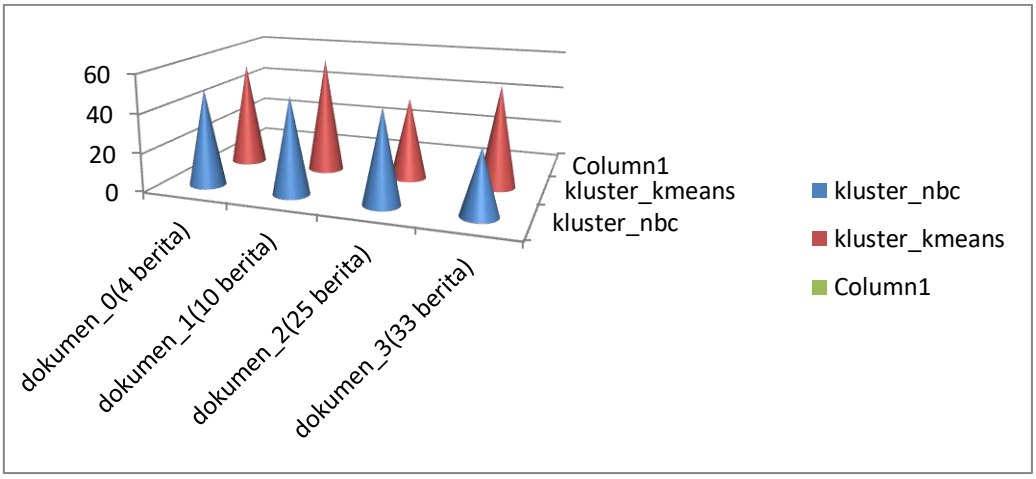
No	Metode	Akurasi ….%				Rata-rata
		dokumen_0 (4 berita)	dokumen_1 (10 berita)	dokumen_2 (25 berita)	dokumen_3 (33 berita)	
1	Naive Bayes Classifier (NBC)	50%	50%	48%	33.33%	45.33%
2	K-Means Clustering					

	Uji_1 (5x)	100% 25% 75% 50% 100%	20% 40% 60% 80% 90%	20% 24% 52% 56% 60%	36.36% 42.42% 51.52% 66.67% 69.7%	
3	Total uji	**70%**	**58%**	**42.4%**	**53.33%**	
	Uji_2 (5x)	0% 75% 25% 50% 25%	50% 40% 80% 90% 40%	56% 24% 60% 20% 52%	69.7% 51.52% 51.52% 42.42% 36.36%	
4	Total uji	**35%**	**60%**	**42.4%**	**50.30%**	
	Uji_3 (5x)	75% 50% 0% 100% 25%	50% 80% 40% 90% 60%	20% 52% 56% 60% 24%	36.36% 42.42% 69.7% 51.52% 66.67%	
5	Total_uji	**50%**	**64%**	**42.4%**	**53.33%**	
	Uji_4 (5x)	50% 25% 50% 0% 100%	40% 90% 50% 60% 80%	56% 24% 60% 20% 56%	69.7% 51.52% 36.36% 66.67% 42.42%	
6	Total uji	**45%**	**64%**	**43.2%**	**53.33%**	**51.38%**
	Uji_5 (5x)	100%	40%	52%	66.67%	

		50% 75% 25% 100%	80% 60% 50% 40%	24% 60% 56% 20%	36.36% 69.7% 51.52% 36.36%	
7	Total_uji	**70%**	**54%**	**42.4%**	**52.12%**	**54.63%**
8	**Total rata-rata uji**	**54%**	**60%**	**42.56%**	**52.48%**	**52.26%**

Keterangan Tabel 4.:

Di dalam tabel analisis akurasi klasifikasi berita terdapat 4 koleksi dokumen berita yaitu: dokumen_0, dokumen_1, dokumen_2, dan dokumen_3. Menggunakan 2 metode *clustering*, yang pertama menggunakan metode *nbc*, hanya diambil 1 sampel hasil akurasi, karena *nbc* prosesnya tidak secara random (acak) sehingga hasil akurasinya tidak akan berubah-ubah. Metode yang kedua adalah *k-means clustering*, dimana masing-masing koleksi dokumen berita dilakukan 5x uji tiap dokumen diambil 5 sampel hasil akurasi dari pengelompokkan berita, karena *k-means clustering* menggunakan proses secara random (acak) sehingga hasilnya akan berubah-ubah. Dari kedua metode yang dilakukan yaitu menggunakan metode *nbc* dan metode *k-means clustering* diperoleh Total Rata-rata metode *nbc* sebesar 45.33 % artinya, dari hasil pengelompokkan berita yang telah dilakukan belum benar karena masih banyak terdapat kesalahan penempatan letak data berita dari hasil *cluster*, hal ini yang menyebabkan rendahnya tingkat akurasi yang diperoleh karena belum mencapai diatas 75 %. Sedangkan hasil akurasi yang diperoleh menggunakan metode *k-means clustering* sebesar 52.26 % jauh lebih tinggi dari perolehan hasil akurasi dengan metode *nbc*. Artinya akurasi sebesar 52.26 % belum benar karena masih terdapat kesalahan letak data berita yang tidak sesuai letaknya. Dari kedua metode yang digunakan dapat disimpulkan bahwa ketepatan data menggunakan dua metode diatas belum benar semua karena masih dibawah 75 %, walaupun metode *k-means clustering* hasilnya lebih tinggi dibandingkan metode *NBC*, dalam ketepatan datanya lebih bagus menggunakan metode nbc karena proses pengelompokannya tidak secara random dan hasilnya tidak akan berubah-ubah.



Gambar IV.13. Grafik akurasi NBC dan k-means clustering

KESIMPULAN

Kesimpulan dari penelitian ini berdasarkan metode yang digunakan adalah ketepatan data menggunakan metode*nbc* diatas belum benar sepenuhnya karena hasil ketepatan data yang diperoleh masih dibawah 75 % dan letak beritanya masih ada yang tidak sesuai letaknya, sedangkan dengan metode *k-means clustering* ketepatan data yang diperoleh sudah tepat karena akurasinya sebesar 100%, sehingga datanya akurat (tepat), walaupun metode *k-means clustering* hasilnya lebih tinggi dibandingkan metode *nbc*, dalam perolehan hasil ketepatan datanya lebih bagus menggunakan metode *nbc* karena proses perhitungan pengelompokan beritanya tidak secara random dan hasilnya tidak akan berubah-ubah.

DAFTAR PUSTAKA

Asroni, R. A. (2015). Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang. *Jurnal Ilmiah Semesta Teknika, Vol, 18, No.1*, Hal 76-82.

Ong, J. O. (2013). Implementasi AlgoritmaK-Means Clustering Untuk Menentukan Strategi Marketing President University. *Jurnal Ilmiah miahTeknik Industri, Vol.12, No.1,*.

Riani. (2016). Klasifikasi Opini Menggunakan Algoritma Naive Bayes Classifier (NBC). *Kerja Praktek, IST AKPRIND YOGYAKARTA*.

Sarwono, Y. T. (n.d.). Aplikasi Model Jaringan Saraf Tiruan Dengan Radial Basis Function Untuk Mendeteksi Kelainan Otak (Stroke Infark). *Jurnal Sistem Informasi*.

W, R. N., Defiyanti, S., & Jajuli, M. (April 2015). Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jurnal Ilmiah Teknologi Informasi Terapan Vol.1, No.2.*

Han, J., Kamber, M., & Pei, J. (2012).*Data Mining: Concepts and Techniques*

*Third Edition.*Waltham: Morgan Kaufmann.

Susanto, Budi. (n.d.).*Text dan Web Mining,Materi Kuliah Teknik Informatika, UKDW Yogyakarta.*