Penerapan Metode *Clustering Text Mining* Untuk Pengelompokan Berita Pada Unstructured Textual *Data*

Nyoman Gede Yudiarta[1], Made Sudarma [2], Wayan Gede Ariastina[3]

***Abstract*—Good governance was a government whose programs were known and beneficial to the people. In Bali Provincial Government which has duty in disseminating information is Bureau of Public Relations Regional Secretariat Bali through media owned. Because at the time of news input to the media in this case Public Relations Bureau website was not included causing the emergence of problems in the form of difficulty knowing the news, which news that goes into certain categories. Clustering was a method to solve the problem. One of the algorithms used in the Clustering method is the K-Means algorithm. This study focused on designing to classify news data into a category using K-Means. To process the documents obtained to make it easier in the process of clustering, was done by preprocess documents first. Document preparation consists of case folding, tokenization, filtering and stemming. Tf-Idf was done to pass the weighting of the *term*s obtained on the preprocessed documents. The results of experiments conducted using different amounts of data that are 50, 100, 200, 300, 400, and 500 data obtained results that the K-Means algorithm applied to cluster news, able to work and provide a satisfactory accuracy, *Precision* average of 70.76% while *Recall* of 70.86% and *Purity* of 0.76 for all test data.**

***Intisari*— Pemerintahan yang baik adalah pemerintahan yang program–programnya diketahui dan bermanfaat bagi masyarakatnya. Pada Pemerintah Provinsi Bali yang memiliki tupoksi dalam melakukan penyebarluasan informasi adalah Biro Humas Setda Provinsi Bali melalui media yang dimiliki. Dikarenakan pada saat input berita ke media dalam hal ini website Biro Humas tidak disertakan kategori menyebabkan timbulnya permasalahan berupa sulitnya mengetahui berita–berita yang mana saja yang masuk ke kategori tertentu. Clustering merupakan metode untuk mengatasi permasalahan tersebut. Salah satu algoritma yang digunakan dalam metode Clustering adalah algoritma K-Means. Penelitian ini berfokus pada perancangan untuk mengelompokan data berita ke suatu kategori dengan menggunakan K-Means. Untuk mengolah dokumen yang didapat agar lebih mempermudah dalam proses clustering, dilakukanlah *preprocessing* dokumen terlebih dahulu. *Preprocessing* dokumen terdiri dari case folding, tokenization, filtering dan stemming. Tf-Idf dilakukan untuk melalukan pembobotan terhadap *term* yang didapatkan pada *preprocessing* dokumen. Hasil coba yang dilakukan dengan menggunakan jumlah data yang berbeda yaitu 50, 100, 200, 300, 400, dan 500 data didapatkan hasil bahwa algoritma K-Means yang diterapkan untuk meng cluster berita, mampu bekerja dan memberikan akurasi yang memuaskan, dengan rata-rata *Precision* sebesar 70,76% sedangkan *Recall* sebesar 70,86% serta *Purity* sebesar 0,76 untuk semua data uji.**

Kata Kunci*— *Clustering, K-Means, Preprocessing, Tf-Idf

I. PENDAHULUAN

Pemerintahan yang baik adalah pemerintahan yang program – programnya diketahui dan bermanfaat bagi masyarakatnya. Masyarakat berhak mengetahui apa saja kegiatan – kegiatan yang telah dilakukan pemerintah untuk memajukan daerahnya. Semua kegiatan dan kebijakan dari mulai rencana kerja sampai hasil dipublikasikan di berbagai media yang dimiliki oleh Pemerintah antara lain melalui *Website*, Sosial Media seperti *Facebook* dan *Twitter*, Pentas Seni, Surat Kabar, dan TV Display.

Di dalam Web Pemerintah terdapat berita – berita yang diumumkan oleh Bagian Publikasi. Dalam hal ini pada saat melakukan input berita, tidak terdapat kategori, penggolongan ataupun pengelompokan jenis berita yang diinputkan sesuai dengan 12 program Aksi Bali Mandara, dengan tidak terdapatnya pengelompokan jenis berita pada saat melakukan input berita menyebabkan timbulnya permasalahan berupa sulitnya mengetahui berita – berita yang mana saja yang masuk ke kategori tertentu sehingga dalam mencari suatu berita dengan topik tertentu memerlukan waktu yang tidak sedikit. Untuk mempermudah dalam pengelolaan berita-berita serta pengelompokan berita yang sesuai dari beberapa dokumen yang melibatkan data text yang tidak terstruktur, maka diperlukan suatu teknik *clustering*. *Clustering* dipakai ketika tidak diketahuinya bagaimana data harus dikelompokkan. *Clustering* dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan atau kemiripan. Sebuah *cluster* adalah sekumpulan objek yang digabung bersama karena persamaan atau kedekatannya [1]. *Clustering* termasuk dalam teknik *unsupervised learning* dimana tidak memerlukan fase training [2].

Karena *clustering text* ini melibatkan data teks yang tidak terstruktur, teknik dalam *text mining* dapat dijadikan sebagai solusi untuk menemukan kata atau pola yang diinginkan untuk dijadikan kunci dalam proses *clustering*. *Text Mining* itu sendiri adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks [3]. Proses pada *text mining* tersebut yang diantaranya adalah *case folding*, *tokenization*, *filtering*, dan *stemming* memiliki tujuan untuk mereduksi atau mengekstrak data serta mengurangi *noise* pada data [4]. Data yang diolah dalam proses *clustering* adalah berupa data berbentuk *vector* maka diperlukan metode pembobotan kata (*term* weighting) untuk menghitung frekuensi kemunculan dari setiap *term*. Dalam hal ini metode yang dipakai dalam pembobotan kata adalah

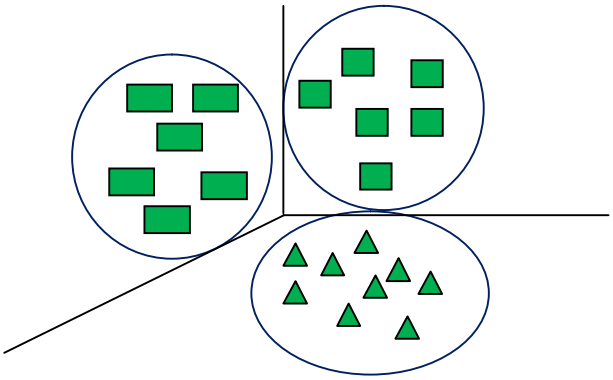
[1]*Mahasiswa, Program Studi Magister Teknik Elektro, Jalan Dewi Supraba VI. No.23 ,Denpasar Bali INDONESIA (tlp:081916153335; e-mail: mankyudiarta@gmail.com)*
[2, 3] *Dosen, Program Studi Magister Teknik Elektro, Jalan Panglima Besar Sudirman, Denpasar Bali Indonesia (tlp: 0361-239599; fax: 0361-239599; e-mail: imasudarma@gmail.com, w.ariastina@gmail.com)*

Metode *Term Frequency (Tf)* dan *Inverse Document Frequency (Idf)*.

Digunakannya Algoritma *K-Means* dalam penelitian ini adalah dikarenakan data inputan yang akan diproses terbilang masih sederhana sehingga lebih cocok menggunakan algoritma *K-Means* serta dilihat dari penelitian sebelumnya dimana Algoritma *K-Means* berhasil melakukan pengelompokan terhadap dokumen teks. Prilianti dan Wijaya [5] melakukan clustering terhadap skripsi mahasiswa di sebuah Universitas Ma Chung, Algoritma *K-Means* berhasil melakukan pengelompokan terhadap dokumen-dokumen srkipsi yang ada dengan nilai *purity* sebesar 76%, artinya sekitar 76% dokumen yang telah diolah telah berhasil dikelompokkan dengan benar oleh sistem. Sistem ini diharapkan dapat mengelompokan berita sesuai dengan persamaan yang dimiliki.

II. METODE PENELITIAN

Clustering merupakan algoritma pengelompokkan sejumlah data menjadi kelompok–kelompok data tertentu (*cluster*) [6], yang bertujuan untuk mengelompokkan data dengan karakteristik yang sama ke suatu wilayah atau kelompok yang sama dan data dengan karakteristik yang berbeda ke wilayah atau kelompok yang lainnya seperti tampak pada Gambar 1:



Gambar 1: Contoh pengelompokan berdasarkan bentuk.

Pada Gambar 1: menunjukkan data dikelompokkan sesuai dengan karakteristik bentuk yang dimiliki dimana data tersebut dikelompokkan menjadi tiga kelompok yaitu kelompok dengan bentuk persegi panjang, kotak dan lingkaran.

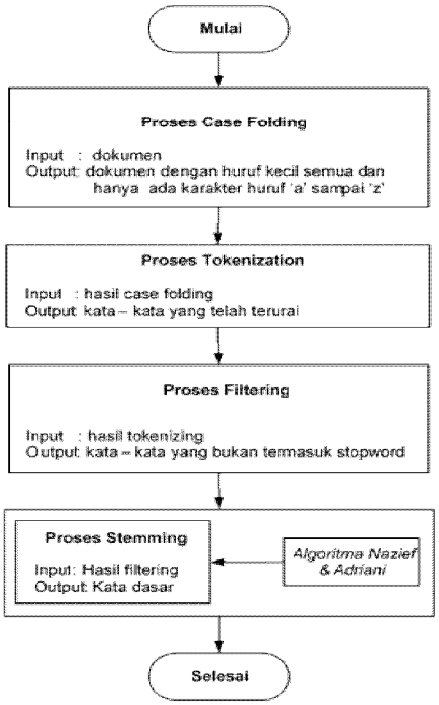
Dalam perancangan penelitian terdapat beberapa tahapan yang digunakan yaitu sebagai berikut:

A. Pengumpulan Data dan Analisis Kebutuhan

Dalam pengumpulan Data peneliti melakukan permintaan permohonan data Pemerintah dengan mengirimkan surat permohonan data. Setelah didapatkannya data tersebut, peneliti melakukan analisa kebutuhan untuk mengidentifikasi jenis informasi apa saja yang diperlukan, evaluasi data dan lingkup laporan yang diinginkan. Sumber data yang akan diolah pada proses *clustering* adalah dari database berita Pemerintah.

B. Preprocessing Dokumen

Untuk mengolah dokumen yang didapat agar lebih mempermudah dalam proses *clustering*, dilakukanlah *preprocessing* dokumen. Preprocessing berfungsi untuk meningkatkan citra, menghilangkan noise, maupun menentukan bagian citra yang akan digunakan dalam tahapan selanjutnya [7]. *Preprocessing* dokumen terdiri dari *case folding* yaitu mengubah semua huruf dalam dokumen menjadi huruf kecil dan karakter selain huruf dihilangkan, selanjutnya *tokenization* atau pemisahan kata, kemudian *filtering* yaitu penghilangan *token* berdasarkan *stopword*, terakhir adalah *stemming* yaitu pencarian kata dasar dari setiap kata. Pada *stemming*, algoritma yang digunakan untuk pencarian kata dasarnya adalah *Algoritma Nazief & Adriani*. Diagram Alur dari *Preprocessing* dapat dilihat pada Gambar 2:



Gambar 2: Diagram alur *preprocessing* dokumen

C. Term Weighting

Setelah dilakukannya *preprocessing* dokumen (*case folding*, *tokenization*, *filtering*, dan *stemming*), dimana Tahapan *preprocessing* akan menghasilkan kumpulan *term* atau kata, selanjutkan dilakukan proses *term weighting* yang nantinya akan diberikan bobot atau nilai dimana bobot tersebut mengindikasikan pentingnya sebuah *term* terhadap dokumen. Penghitungan bobot tiap *term* dicari pada setiap dokumen bertujuan untuk dapat mengetahui ketersediaan dan kemiripan suatu *term* di dalam dokumen [8]. Semakin banyak *term* tersebut muncul pada koleksi dokumen, semakin tinggi nilai atau bobot *term* tersebut. Setelah tahapan pemberian bobot selesai barulah dilanjutkan ke proses *clustering*. Dalam *Term Weighting*, metode yang digunakan dalam melakukan pembobotan adalah metode *Tf-Idf*.

Term frequency (*Tf*) adalah algoritma pembobotan heuristik yang menentukan bobot dokumen berdasarkan kemunculan *term* [9]. Terdapat empat buah algoritma *TF* yaitu *Raw TF*, *Logarithmic TF*, *Binary TF*, dan *Augmented TF* [9], dalam hal ini yang digunakan adalaha *Raw Tf*. *Raw Tf*

merupakan penentuan bobot suatu dokumen terhadap istilah dengan menghitung frekuensi kemunculan suatu istilah tersebut pada dokumen. Semakin sering sebuah istilah/kata itu muncul, semakin tinggi bobot dokumen untuk istilah/kata tersebut, dan juga sebaliknya.

Inverse Document Frequency (Idf) fokus pada kemunculan *term* pada keseluruhan koleksi teks. Pada *Idf*, *term* yang jarang muncul pada keseluruhan koleksi *term* dinilai lebih berharga. *Inverse Document Frequency (Idf)* dihitung dengan menggunakan formula (1).

$$Idf = \log\left(\frac{jumlah\ seluruh\ dokumen\ dalam\ koleksi}{jumlah\ dokumen\ yang\ mengandung\ istilah}\right) \quad (1)$$

Dengan demikian rumus umum untuk perhitungan *Tf-Idf* adalah penggabungan dari formula perhitungan *Raw Tf* dengan formula *Idf* dengan cara mengalikan nilai *Term Frequency* (*Tf*) dengan nilai *Inverse Document Frequency* (*Idf*).

D. Proses Clustering

Pada proses ini dilakukan pengelompokan berita secara otomatis. Setelah dilakukannya *preprocessing* dokumen yang menghasilkan kata atau *term* pada setiap dokumen. Selanjutnya dilakukan *Term Weighting* untuk membobotan setiap *term* tersebut, dimana nantinya hasil perhitungan dari *Tf-Idf* dibentuk suatu *vektor*. Setelah mendapatkan *vektor* tersebut dilanjutkan dengan proses *clustering* dengan menggunakan algoritma *K-Means*, dengan langkah sebagai berikut :

1. Menentukan banyaknya kelompok, dimana kelompok telah ditentukan sebanyak 12 kelompok, yaitu dari 12 program Aksi Bali Mandara.
2. Kemudian objek *vektor* yang telah didapatkan dari proses pembobotan dialokasikan, dan selanjutnya menentukan *centroid* nya secara random,
3. Setelah *centroid* ditentukan maka selanjutnya menghitung jarak antara 2 *vektor*, dalam hal ini adalah jarak antara *centroid* dengan objek atau *term* dengan menggunakan metode *Euclidean Distance*. Adapun rumus yang digunakan untuk menghitung jarak antara 2 vektor dengan *Euclidean Distance* [10], seperti dalam formula (2).

$$\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2 \quad (2)$$

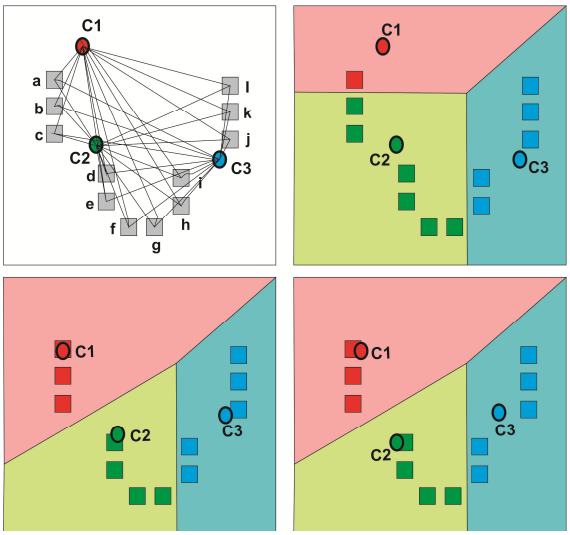
dengan :
dij = tingkat perbedaan
n = jumlah vektor
xik = vektor citra input
xjk = vektor citra pembanding / output

4. Jika centroid berubah lagi proses kembali ke langkah 3 dengan penentuan posisi centroid baru dengan menggunakan persamaan (3).

$$\boldsymbol{v} = \frac{\sum_{i=1}^{n} x_i}{n};\ i = 1,2,3,\ldots n \quad (3)$$

dengan :
v = centroid pada cluster
xi = objek ke-i
n = banyaknya objek/jumlah objek yang menjadi anggota cluster

jika posisi *centroid* tidak berubah lagi berarti proses *clustering* selesai dan hasil yang didapat adalah pengelompokkan objek dalam kategori tertentu berdasarkan centroid yang terdekat. Berikut ini Ilustrasi dalam proses pembentukan anggota suatu clustering dengan menggunakan Algoritma *K-Mean*, seperti pada Gambar 3:



Gambar 3: Ilustrasi Pembentukan Anggota Clustering dengan K-Means

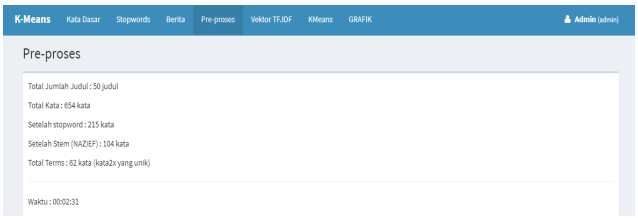
E. Evaluasi

Setelah melakukan pembuatan sistem, dilakukan evaluasi terhadap hasil dari clustering dengan melakukan perhitungan terhadap *Precision*, *Recall*, dan *Purity* dari *cluster* yang dihasilkan, apakah sudah tepat kelompok *cluster* yang dibentuk oleh sistem. Sebelum melakukan penghitungan nilai dari precesion, *recall*, dan *purity* akan dilakukan pelabelan manual. Tujuan dilakukannya pelabelan manual adalah sebagai bahan perbandingan untuk perhitungan hasil cluster yang dilakukan oleh sistem, dimana pelabelan manual sebelumnya dilakukan oleh pakar yang membidangi informasi program-program Pemerintah.

III. HASIL DAN EVALUASI

3.1 *HASIL*

Tahapan pertama dalam penelitian ini yang harus dilewati adalah tahapan *preprocessing* dokumen. Pada penelitian ini akan dilakukan perhitungan terhadap 50 data percobaan. Pada 50 judul berita yang di proses, terdapat 654 kata, setelah dilakukan tahapan *preprocessing* dokumen berupa *case folding*, *tokenization*, *filtering*, dan terakhir *stemming* total kata menjadi 104 kata dengan total *term* yang akan diproses lebih lanjut adalah sebanyak 62 kata-kata yang unik. Hasil sistem dapat dilihat pada Gambar 4:



Gambar 3: Hasil *Preprocessing* Sistem

berikut *term* yang didapatkan, setelah hasil *preprocessing* yang dilakukan oleh sistem adalah pada Tabel 1.

TABEL I
HASIL *TERM* DARI *PREPROCESSING*

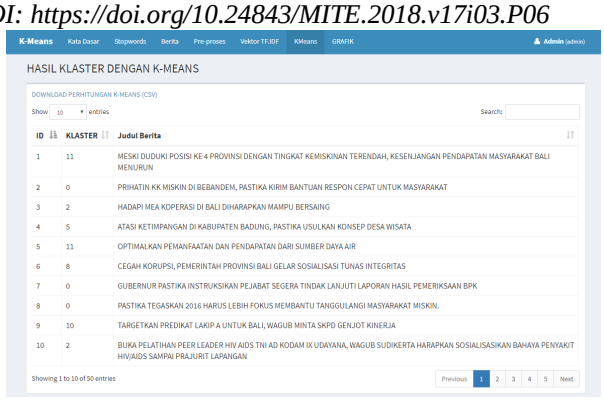
No	*Term*	No	*Term*
1	miskin	32	pramuka
2	pendapatan	33	bina
3	mea	34	sadar
4	koperasi	35	muda
5	timpang	36	jamin
6	desa	37	akses
7	wisata	38	simakrama
8	sumber	39	pad
9	daya	40	bupati
10	perintah	41	anak
11	lapor	42	tonggak
12	periksa	43	puri
13	bpk	44	krama
14	kinerja	45	ibu
15	sakit	46	komponen
16	rumah	47	tani
17	pariwisata	48	pohon
18	budaya	49	studi
19	distribusi	50	rekomendasi
20	akuntabel	51	sosialisasi
21	disdikpora	52	didik
22	pns	53	olahraga
23	disiplin	54	generasi
24	siwaratri	55	lintas
25	besakih	56	agama
26	umat	57	akuntansi
27	tuhan	58	akrual
28	deklarasi	59	jurnalistik
29	mental	60	berita
30	reformasi	61	giat
31	birokrasi	62	sejahtera

Tahapan selanjutnya adalah proses prekalian antara *Term Frequency* (*Tf*) dengan *Inverse Document Frequency* (*Idf*). Berikut hasil proses *Tf*Idf* dapat dilihat pada Tabel 2.

TABEL 2
HASIL *TF*IDF*

Doc	Miskin	Pendapatan	mea	koperasi	s/d …	sejahtera
1	0.796	1.398	0	0	…	0
2	0.796	0	0	0	…	0
3	0	0	1.699	1.398	…	0
4	0	0	0	0	…	0
5	0	1.398	0	0	…	0
6	0	0	0	0	…	0
7	0	0	0	0	…	0
8	0.796	0	0	0	…	0
9	0	0	0	0	…	0
10	0	0	0	0	…	0
11	0	0	0	0	…	0
12	0.796	0	0	0	…	0
14	0	0	0	0	…	0
15	0	0	0	0	…	0
16	0.796	0	0	0	…	0
17	0	0	0	0	…	0
18	0	0	0	0	…	0
19	0	0	0	0	…	0
20	0	0	0	0	…	0
21	0	0	0	0	…	0
22	0	0	0	0	…	0
23	0	0	0	0	…	0
24	0	0	0	0	…	0
25	0	0	0	0	…	0
26	0	0	0	0	…	0
27	0	0	0	1.398	…	0
28	0	0	0	0	…	0
29	0	0	0	0	…	0
30	0.796	0	0	0	…	0
31	0	0	0	0	…	0
32	0	0	0	0	…	0
33	0	0	0	0	…	0
34	0	0	0	0	…	0
35	0.796	0	0	0	…	0
36	0	0	0	0	…	0
37	0	0	0	0	…	0
38	0	0	0	0	…	0
39	0	0	0	0	…	0
40	0	0	0	0	…	0
41	0	0	0	0	…	0
42	0	0	0	0	…	0
43	0	0	0	0	…	0
44	0	0	0	0	…	0
45	0	0	0	0	…	0
46	0	0	0	0	…	0
47	0	0	0	0	…	0
48	0	0	0	0	…	0
49	0	0	0	0	…	0
50	0.796	0	0	0	…	1.699

Setelah proses *preprocessing* dukumen dan mengubah *term* menjadi data *vektor* melalui perkalian *Tf*Idf*, maka selanjutnya dilakukan proses peng*clusteran* dengan menggunakan *K-Means*. Berikut adalah hasil dari *Clustering* sistem dengan menggunakan *K-Means* dapat dilihat pada Gambar 4.



Gambar 3: Hasil *Preprocessing* Sistem

3.2 Evaluasi dengan Metode Precision, Recall, dan Purity

Pengujian *precision, recall* dan *purity* dilakukan untuk mengetahui tingkat akurasi dari hasil *clustering* yang didapatkan oleh sistem. Data yang akan di evaluasi adalah hasil *clustering* dengan jumlah data : 50, 100, 200, 300, 400, dan 500. Digunakannya data tersebut untuk melihat perbandingan dari hasil clustering yang didapat oleh sistem pada jumlah data yang berbeda. Sebelum masuk ke pengujian menggunakan metode *precision, recall* dan *purity* terlebih dahulu dilakukan pemberian label manual oleh Pemerintah. Berikut perbandingan pelabelan pada pengujian 50 data yang dilakukan oleh Pemerintah dengan *cluste*r yang dilakukan oleh sistem jika dikelompokan sesuai *cluster* nya, maka didapatkan hasil seperti pada Tabel 3.

TABEL 3
PENGELOMPOKAN HASIL CLUSTER 50 DATA

Cluster	**Banyaknya Judul Berita**	**Kategori**
Cluster 0	8 judul berita	Bidang Sosial
Cluster 1	2 judul berita	Bidang Perekonomian
Cluster 2	9 judul berita	Bidang Kesehatan
Cluster 3	1 judul berita	Bidang Lingkungan dan Pertanian
Cluster 4	4 judul berita	Bidang Seni Budaya dan Pariwisata
Cluster 5	3 judul berita	Bidang Pendidikan
Cluster 6	4 judul berita	Bidang Pemuda dan Olah Raga
Cluster 7	3 judul berita	Bidang Demokrasi dan HAM
Cluster 8	10 judul berita	Bidang Keamanan dan Ketertiban Masyarakat
Cluster 9	1 judul berita	Bidang Infrastruktur
Cluster 10	2 judul berita	Bidang Pemberdayaan Perempuan
Cluster 11	3 judul berita	Bidang Ekonomi Kerakyatan dan Ketenagakerjaan

Dalam hal ini kategori dari setiap cluster akan ditentukan oleh peneliti, dikarenakan sistem tidak mengetahui kategori dari setiap *cluster*, sistem hanya melakukan peng*clusteran* dari setiap judul yang telah ditentukan. Untuk mempermudah dalam menghitung nilai *precision* dan *recall* maka akan ditelusuri data yang relevan dan yang tidak relevan pada data yang telah di *cluster*. Dapat dilihat pada tabel 4.

TABEL 4
HASIL PENELUSURAN *CLUSTER* 50 DATA

Cluster	**Rele Van (a)**	**Tidak Relevan (b)**	**Ditemukan (a+b)**	**Tidak Ditemukan (d)**	**Total Relevan dalam Koleksi (a+d)**
Cluster 0	7	1	8	0	7
Cluster 1	2	0	2	6	8
Cluster 2	8	1	9	0	8
Cluster 3	1	0	1	0	1
Cluster 4	4	0	4	1	5
Cluster 5	1	2	3	1	2
Cluster 6	2	2	4	2	4
Cluster 7	3	0	3	0	3
Cluster 8	8	2	10	0	8
Cluster 9	1	0	1	0	1
Cluster 10	1	1	2	0	1
Cluster 11	1	2	3	1	2
Total	39	11	50	11	50

Berikut nilai *Precision* dan *Recall* dari masing masing *cluster* dapat dilihat pada Tabel 5 dan Tabel 6.

TABEL 5
HASIL *PRECISION* 50 DATA

Cluster	**Kategori**	***Precision***
Cluster 0	Bidang Sosial	87.50 %
Cluster 1	Bidang Perekonomian	100.00 %
Cluster 2	Bidang Kesehatan	88.89 %
Cluster 3	Bidang Lingkungan dan Pertanian	100.00 %
Cluster 4	Bidang Seni Budaya dan Pariwisata	100.00 %
Cluster 5	Bidang Pendidikan	33.33 %
Cluster 6	Bidang Pemuda dan Olah Raga	50.00 %
Cluster 7	Bidang Demokrasi dan HAM	100.00 %
Cluster 8	Bidang Keamanan dan Ketertiban Masyarakat	80.00 %
Cluster 9	Bidang Infrastruktur	100.00 %
Cluster 10	Bidang Pemberdayaan Perempuan	50.00 %
Cluster 11	Bidang Ekonomi Kerakyatan dan Ketenagakerjaan	33.33 %
	Rata-rata	76.92%

Dilihat dari tabel diatas, bahwa sebagian besar setiap cluster memiliki presisi yang bagus, artinya pada satu cluster, data benar yang didapat lebih banyak daripada data yang salah, hanya beberapa cluster yang memiliki hasil presisi yang kurang yaitu cluster 5 dan cluster 11.

TABEL 6
HASIL *RECALL* 50 DATA

Cluster	**Kategori**	***Recall***
Cluster 0	Bidang Sosial	100.00 %
Cluster 1	Bidang Perekonomian	25.00 %
Cluster 2	Bidang Kesehatan	100.00 %
Cluster 3	Bidang Lingkungan dan Pertanian	100.00 %
Cluster 4	Bidang Seni Budaya dan Pariwisata	80.00 %
Cluster 5	Bidang Pendidikan	50.00 %
Cluster 6	Bidang Pemuda dan Olah Raga	50.00 %
Cluster 7	Bidang Demokrasi dan HAM	100.00 %

Cluster 8	Bidang Keamanan dan Ketertiban Masyarakat	100.00 %
Cluster 9	Bidang Infrastruktur	100.00 %
Cluster 10	Bidang Pemberdayaan Perempuan	100.00 %
Cluster 11	Bidang Ekonomi Kerakyatan dan Ketenagakerjaan	50.00 %
	Rata-rata	79.58%

Pada Tabel 6. dapat dilihat bahwa dalam hal ini sistem berhasil mengelompokan judul berita ke dalam kelompok yang tepat, hanya 1 *cluster* yang kurang ditemukan sesuai dengan total yang harus ditemukan yaitu pada cluster 1.

Berikut hasil perbandingan nilai rata-rata dari *precision* dan *recall* serta nilai *purity* dari 50 data, 100 data, 200 data, 300 data, 400 data, dan 500 data dapat dilihat pada Tabel 7.

TABEL 7
HASIL PERBANDINGAN NILAI PRECISION, *RECALL, DAN PURITY*

No	Data	*Precision*	*Recall*	*Purity*
1	*50*	76,92 %	79,58 %	0,78
2	100	71,84 %	72,38 %	0,79
3	200	71,17 %	63,26 %	0,78
4	300	76,44 %	70,73 %	0,83
5	400	72,27 %	65,92 %	0,81
6	500	70.00 %	66,03 %	0,79

Pada Tabel 7. dapat dilihat bahwa pada pengujian 50 data memiliki tingkat rata-rata *precision* dan *recall* paling tinggi yaitu 76,92% untuk *precision* nya sedangkan untuk *recall* nya sebesar 79,58% dari pengujian data yang lainnya. ini berarti bahwa penempatan data pada setiap cluster nya di pengujian 50 data kebanyakan sudah tepat. Sedangkan untuk *Purity* nya nilai yang paling tinggi adalah pada pengujian 300 data yaitu sebesar 0,83.

IV. KESIMPULAN

Kesimpulan yang dapat ditarik dari penelitian ini adalah sebagai berikut :

1. Penentuan *centroid* awal (titik pusat) pada Algoritma *K-Means* sangat berpengaruh pada hasil *cluster*, dimana *centroid* awal tersebut ditentukan secara acak, sehingga terkadang tingkat keakuratannya kurang baik, maka dari itu perlu dilakukan proses uji coba berkali kali agar mendapatkan hasil *cluster* yang baik.
2. Pada beberapa kelompok data yang diuji, pengujian 50 data memiliki rata – rata persentase nilai *Precision* dan *Recall* yang paling besar yaitu 76,92% untuk *precision* dan sebesar 79,58% untuk *recall* nya. Sedang kan untuk nilai *purity* nya yang terbesar terdapat pada pengujian 300 data yaitu sebesar 0,83. Dengan demikian dapat dikatakan bahwa Algoritma *K-Means* mampu mengelompokkan dokumen ke dalam 12 kelompok, serta melakukan pengelompokan dokumen dalam jumlah yang banyak.

REFERENSI

[1] Herny Februariyanti Dan Dwi Budi Santoso, 2017, "Hierarchical Agglomerative Clustering Untuk Pengelompokan Skripsi Mahasiswa," Prosiding SINITAK 2017, ISBN: 978-602-8557-20-7.
[2] Pivin Suwrmayanti, I Ketut Gede Darma Putra, I Nyoman Satya Kumara, "Optimasi Pusat Cluster K-Prototype dengan Algoritma Genetika," Teknologi Elektro, Vol. 13 No. 2 Juli-Desember 2014.
[3] PenambanganTeks, https://id.wikipedia.org/ (diakses tanggal 27 Juni 2015).
[4] Thopo Martha Akbar, Angelina Prima Kurniati, Moch Arif Bijaksana, 2012 "Analisis Perbandingan Metode Pembobotan Kata Tf.Idf Dan Tf.Rf Terhadap Performansi Kategorisasi Teks".
[5] Kestrilia Rega Prilianti, Hendra Wijaya, 2014, "Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik Skripsi Dengan Metode K-Means Clustering," Jurnal Cybermatika, Vol. 2 No. 1.
[6] Mardiani, 2014, " Perbandingan Algoritma K-Means dan EM untuk Clusterisasi Nilai Mahasiswa Berdasarkan Asal Sekolah," Citec Journal, Vol. 1, No. 4, ISSN: 2354-5771.
[7] Ni Putu Sutramiani, I Ketut Gede Darma Putra, Made Sudarma, "Local Adaptive Thresholding pada Preprocessing Citra Lontar Aksara Bali," Jurnal Teknologi Elektro, Vol.14, No.1,Januari-Juni 2015.
[8] Pausta Yugianus, Harry Soekotjo Dachlan, dan Rini Nur Hasanah, 2013 "Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback", EECCIS Vol. 7, No. 1, Juni 2013.
[9] Sendhy Rachmat Wurdianarto, Sendi Novianto, Umi Rosyidah, 2014, Perbandingan Euclidean Distance Dengan Canberra Distance Pada Face Recognition, Techno.COM, Vol. 13, No. 1 : 31-37
[10] Ediyanto, Muhlasah Novitasari Mara, Neva Satyahadewi, 2013, "Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis," Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) Volume 02 , No. 2, Hal 133 – 136