PERANCANGAN SISTEM PENDETEKSI PLAGIARISME TERHADAP TOPIK PENELITIAN MENGGUNAKAN METODE *K-MEANS CLUSTERING* DAN MODEL *BAYESIAN*

Mufti Ari Bianto¹⁾, Siti Rahayu²⁾, Miftahul Huda³⁾, Kusrini⁴⁾

1), 2),3),4) Magister Teknik Informatika Universitas AMIKOM Yogyakarta
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
Email: muftyari@gmail.com^{1),} sitirahayufikom@gmail.com^{2),} miftahulh2@gmail.com^{3),} kusrini@amikom.ac.id⁴⁾

Abstrak

Ancaman dalam dunia akademik khususnya pada bidang penulisan merupakan penjiplakan atau yang bisa kita sebuat sebagai plagiarisme. Menjiplak karya orang lain kemudian mengakuinya sebagai karya pribadi maka hal itu merupakan bagian dari tindakan plagiarisme. Adanya sebuah sistem pendeteksi plagiarisme pada sebuah dokumen teks biasanya menggunakan suatu algoritma yang kemudian akan diimplementasikan untuk mencari kesamaan kata atau kalimat pada setiap dokumen yang berbeda.

K-means merupakan salah satu metode clustering yang dapat mempartisi data yang ada ke dalam bentuk satu atau lebih cluster. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam cluster yang lain untuk kemudian akan dicari nilai terdekat menggunakan metode bayesian sehingga akan ditemukan nilai terdekat dari kemiripan antar dokumen. Perancangan dilakukan dengan Sampel dokumen training sebanyak 10 dokumen dan 1 Dokumen Query. Perhitungan dihitung menggunakan Microsoft Excel. Dokumen yang berhasil dicluster paling dekat dengan dokumen Query kemudian diuji kemiripannya berdasarkan kesamaan pola kalimat dan kesamaan term. Hasil yang diperoleh menggunakan kesamaan term menemukan lebih banyak dokumen dengan presentase kemiripan yaitu 66%-100%, sedangkan berdasarkan kesamaan pola hanya menemukan 1 dokumen dengan kemiripan 100%.

Keywords: plagiarisme, kmeans clustering, bayesian.

1. Pendahuluan

Plagiarisme selalu menjadi sorotan di bidang akademis, karena hal ini dapat didefinisikan sebagai pencurian hasil karya orang lain bahkan sampai mengaku sebagai karya yang telah di buatnya sendiri [1]. Ketika penyalinan hasil karya orang lain tersebut di sertai nama pengarang beserta judul refrensi maka hal tersebut bukanlah plagiarisme. Namun faktanya menyalin suatu hasil pekerjaan orang lain sebagaian maupun keseluruhan

tanpa menyertakan referensi ke sumber aslinya sudah menjadi hal biasa yang sering ditemukan pada suatu penelitian.

ISSN: 2302-3805

Plagiarisme dapat dicegah sedini mungkin dan diminimalisir dengan cara membandingkan manual kedua jenis dokumen tersebut, tetapi cara tersebut tentu kurang efektif [2]. Pencegahan plagiarisme bisa dilakukan melalui hukum dan diperlukan alat atau sebuah sistem pendeteksi plagiarisme terhadap topik penelitian untuk mengetahui plagiarisme secara otomatis dan akurat [3].

Perancangan sistem merupakan suatu proses pengembangan spesifikasi sistem baru berdasarkan hasil rekomendasi analisis sistem [4]. Plagiarisme merupakan bahasa latin dari plagiarus yang memiliki makna penculik [5]. Menurut [6] menetapkan tujuh ciri-ciri tindakan plagiarisme sebagai berikut:

- 1. karya kelompok sebagai kepunyaan atau hasil sendiri.
- 2. menyajikan tulisan yang sama dalam kesempatan yang berbeda tanpa mengakui tulisan sendiri.
- 3. mengakui gagasan orang lain sebagai pemikiran sendiri.
- 4. mengakui teman orang lain sebagai kepunyaan sendiri.
- 5. mengakui menyebutkan asal-usulnya
- 6. meringkas dan memparafrasekan (mengutip tak langsung) tanpa menyebutkan sumbernya.
- 7. meringkas dan memparafrasekan dengan menyebut sumbernya, tetapi rangkaian kalimat dan pilihan katanya masih terlalu sama dengan sumbernya.

Penelitian yang dilakukan oleh [7] menyebutkan bahwa walaupun metode *K-means clustering* hasilnya lebih tinggi dibandingkan metode *naive bayes*, dalam perolehan hasil ketepatan datanya lebih bagus menggunakan metode *naive bayes* karena proses perhitungan pengelompokan beritanya tidak secara random dan hasilnya tidak akan berubah-ubah. Oleh karena itu diperlukan lebih banyak pengimplementasian suatu sistem pendeteksi plagiarisme terhadap topik penelitian, sehingga akan membantu meminimalisir plagiarisme pada topik penelitian selanjutnya.

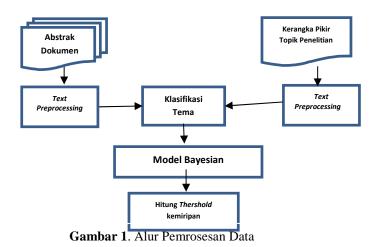
UNIVERSITAS AMIKOM Yogyakarta, 10 Februari 2018

Clustering merupakan teknik pada data mining yang digunakan untuk menganalisis data dan memecahkan permasalahan dengan cara mengelompokkan data atau membagi dataset kedalam partisi [8]. Dalam hal ini ada beberapa algoritma yang sering di gunakan dalam clustering data, salah satunya adalah algoritma K-Means clustering. K-means Clustering adalah algoritma clusterisasi yang berfungsi untuk mengelompokkan data berdasarkan titik pusat (centroid) yang terdekat dengan data [9].

Meskipun pada algoritma ini memiliki nilai keakuratan yang tinggi, namun hasilnya selalu berubah-ubah dikarenakan pengelompokan data yang ada dilakukan secara random. Maka dalam penelitian ini penulis memadukan antara metode *k-means clustering* dengan metode *naive bayes*. Metode *Naive Bayes* adalah metode yang digunakan untuk memprediksi suatu probabilitas. Klasifikasi *Bayes* adalah klasifikasi yang bersifat statistik yang digunakan untuk memprediksi kelas suatu anggota probabilitas [10]. Konsep ini digunakan untuk mecari kedekatan, kemiripan dan hubungan keterkaitan dalam sebuah dokumen text yaitu tentang topik penelitian.

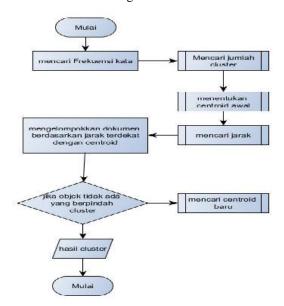
2. Pembahasan

Prosedur kerja dari rencana aplikasi adalah pada **Gambar 1**:



Input pada aplikasi ini berupa Abstrak Dokumen karya ilmiah (tugas akhir, jurnal publikasi, tesis) sebagai data training dan kerangka pikir topik penelitian sebagai *query* yang diinput untuk diperiksa kemungkinan plagiasinya.

a. Desain & Analisis Algoritma



Gambar 2. Alur Proses K-Means

Sistem akan mulai melakukan klasifikasi terhadap setiap dokumen training yang telah tersimpan dalam folder/database. Seperti pada **Gambar 2**, dokumen terlebih dahulu melalui proses text preprocessing untuk di dapat bobot tiap kata kunci yang terkandung pada tiap dokumen.

Kemudian menentukan bobot *term* dari dokumen *query* sebagai titik *centroid* awal untuk mencari dokumen dengan jarak terdekat dengan dokumen *query*.

Tahap K-means secara umum adalah:

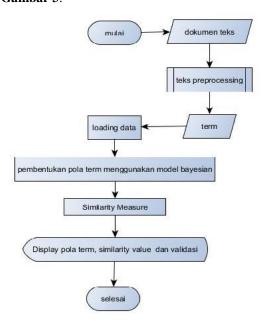
- 1. Menentukan banyaknya cluster K
- 2. Menentukan centroid
- 3. Apakah *centroid*-nya berubah? Jika ya, hitung jarak data dari *centroid*, jika tidak selesai.
- 4. Mengelompokkan data berdasarkan jarak yang terdekat.

Jarak (euclidean) dihitung dengan rumus :

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n} (xi - yi)^{2}}.....(1)$$

Pada perancangan ini, banyaknya *cluster* K ditentukan sebanyak **K=3**. Dengan variabel *centroid* sebanyak *term* yang terdapat pada *query*. Nilai *centroid* untuk *cluster* pertama akan diambil dari nilai TF-IDF dokumen *query*, dan dua *cluster* lainnya akan diambil secara random dari dokumen *training*. Dengan demikian dokumen yang terdapat pada *cluster* 1 yang akan digunakan untuk diperhitungkan kemiripannya dengan dokumen *query*.

Dokumen dengan nilai jarak terdekat dengan *query* (*cluster* 1) kemudian diuji kemiripan kontennya dengan Model *Bayesian* dengan alur seperti pada **Gambar 3**.



Gambar 3. Alur Proses Model Bayesian

Dokumen-dokumen yang masuk pada *cluster* yang sama dengan dokumen *query* kemudian di *loading* isi datanya untuk diperoleh data pola kalimat yang terdapat dalam dokumen tersebut.

Dokumen dipisah menjadi kalimat dan tiap kata pada kalimat tersebut diubah menjadi nomor urut *term* yang telah di dapat melalui teks *preprocessing* dari semua dokumen terkait.

Setelah pola kalimat dari tiap dokumen terbentuk kemudian dilakukan perhitungan similaritas antara dokumen *query* dengan dokumen pembanding dengan rumus *similarity* sebagai berikut:

$$similarity(doci, docj) = \frac{jml_tb(doci, docj)}{tot_term(doci, docj)} \times 100\%...(2)$$

Dengan,

Jml_tb = Jumlah *term* berurutan yang ada pada dokumen-i dan dokumen-j Tot_term = total *term* pada dokumen-i dan dokumen-j

ukuran dokumen dianggap terdapat plagiasi apabila kemiripan mencapai *threshold* >=50% untuk penemuan pola kalimat yang sama. Seperti yang disampaikan [6] bahwa rangkaian kalimat dan pilihan katanya masih terlalu sama dengan sumbernya merupakan ciri dari plagiasi. Untuk itu pada pengujian hasil didasarkan pada presentase kemiripan pola kalimat dan kesamaan kata kunci.

3. Rancangan Hasil

Tahap perancangan hasil dilakukan menggunakan Microsoft Excel. Pada **Gambar 4.** hasil perhitungan bobot tiap kata kunci setiap dokumen yang berelasi dengan kata kunci dokumen *query*. Kata kunci yang tidak berkesesuaian dengan kata kunci *query* tidak akan digunakan pada tahap ini, Konsep ini dilakukan untuk memudahkan proses *clustering* dokumen training yang terdekat dengan dokumen *query*.

	Q	0.6655	0.3979	0.2218	0.2218	0.2218	0.2218	0.2218	0.3979
No	Data Dok	Dokumen	Sedia	Digital	Kait	jiplak	cantum	sumber	referensi
1	D1	0.6655	0.3979	0.2218	0.2218	0.2218	0.2218	0.2218	0.3979
2	D2	0	0	0	0	0	0	0	0
3	D3	0	0	0	0.22185	0	0	0	0
4	D4	0.22185	0	0.22185	0	0.22184875	0.22185	0.22184875	0
5	05	0	0	0	0.22185	0	0	0	0
5	D6	0	0	0	0.22185	0	0	0	0
7	D7	0.6655	0.3979	0.2218	0.2218	0	0	. 0	0
3	D8	0.22185	0	0	0.22185	0.2218	0	0	0
9	D9	0	0	0	0.22185	0	0	0	0
10	D10	0	0.22184875	0.22185	0.22185	0	0.22135	0	0

Gambar 4. Hasil TF.IDF dokumen

Dari 3 *cluster* yang ditentukan , masing-masing ditentukan nilai *centroid* sebagai berikut :

C1=Q={0.6555,0.3979,0.2218,0.2218,0.2218, 0.3979}

C2=D5={0,0,0,0.22185,0,0,0,0}

C3=D8={0,0.2218,0.2218,0.2218,0,0.2218,0,0}

Setelah menetukan nilai *centroid*, tahap selanjutnya adalah menghitung jarak tiap dokumen training dengan *centoid*. Contoh:

Jarak(D3,Q)=| 0- 0.6555| + | 0- 0.3979| + |0-

0.2218| + |0.2218-0.2218| + |0-0.2218| + |0-0.2218| + |0-0.2218| + |0-0.3979| =**2.3488**

Sehingga diperoleh data jarak seperti pada **Gambar 5**:

	Q						
No	Data Dok	DISTANCE C1	DISTANCE C2	DISTANCE C3	C1	C2	C3
1	D1	0.0000	2.1270	1.9051	1		
2	D2	2.5707	0.2218	0.6555		1	
3	D3	2.3488	0.0000	0.4437		1	
4	D4	1.4514	1,3311	0.8874			1
5	D5	2.3488	0.0000	0.4437		1	
6	D6	2.3488	0.0000	0.4437		1	
7	D7	1.0535	1.2853	1.2853	1		
8	D8	1.9051	0.4437	0.0000			1
9	D9	2.3488	0.0000	0.4437		1	
10	D10	1.6833	0.5655	1.1092		1	

Gambar 5. Hasil perhitungan jarak tiap dokumen dengan *centroid* tiap *cluster*

Setelah melakukan perhitungan jarak , maka tiap dokumen kemudian terbagi kebebrapa *cluster* yang telah ditentukan. Hitung kembali jarak dengan nilai *centroid* baru untuk memastikan apakah ada dokumen yang berpindah *cluster* atau tidak sehingga menghasilkan *centroid* baru speerti pada **Gambar 6**.

Jika sudah tidak ada lagi dokumen yang berpindah cluster maka pengelompokan dokumen telah selesai.

										1
		0.5655	0.3979	0.2213	0.2218	0.2218	0.2218	0.2218	0.3979	l
	C1	0.66555	0.39794	0.22185	0.22185	0	0	0	0	l
		0.5655	0.3979	0.2218	0.2218	0.1109	0.1109	0.1109	0.1990	
1		0	0	0	0	0	0	0	0	1
		0	0	0	0.22185	0	0	0	0	
	C2	0	0	0	0.22185	0	0	0	0	
		0	0	0	0.22185	0	0	0	0	l
		0	0	0	0.16639	0	0	0	0	
		0.22185	0	0.22185	0	0.22185	0.22185	0.2218	0	Ī
	С3	0.22185	0	0	0.22185	0.22185	0	0	0	
		0.22185	0	0.11092	0.11092	0.22185	0.11092	0.1109	0	

Gambar 6. Perhitungan titik centroid baru untuk tiap cluster

Dari nilai centroid baru yang diperoleh kemudian dihitung kembali jarak anatara dokumen dengan nilai centroid tiap cluster dan diperoleh hasil cluster seperti Gambar 7:

DISTANCE C1	DISTANCE C2	DISTANCE C3	C1	C2	C3
0.5317	2.2934	1.6833	1		
2.0389	0.1664	0.8874		1	
1.8171	0.0555	0.8874	- 0	1	
1.5952	1.2756	0.4437	- 3		1
1.8171	0.0555	0.8874	- 9	1	
1.8171	0.0555	0.8874		1	
0.5317	1.3408	1.5072	1		
1.5952	0.4992	0.0000	3		1
1.8171	0.0555	0.8874	- 8	1	
1.3734	0.7210	1.1092		1	

Gambar 7. Hasil Cluster untuk centroid baru (iterasi

Dari hasil perhitungan jarak dengan centroid baru, tidak ada dokumen yang berpindah cluster maka proses

telah Maka data sebagai **Tabel**

hasil

C1	C2	C3
D1	D2	D4
D7	D3	D8
	D5	
	D6	
	D9	
	D10	

klasifikasi selesai. diperoleh cluster berikut: 1. Dokumen cluster

Maka dokumen yang terdekat dengan query adalah dokumen 1 dan dokumen 2, kedua dokumen ini yang kemudian diproses untuk tahap selanjutnya dalam

No	Dok	isi term	Pola term	Total term
1	D1	dokumen sedia digital. Kait jiplak cantum sumber referensi.	T1 T2 T3. T4 T5 T6 T7 T8.	8
2	D7	dokumen sedia digital kait	T1 T2 T3 T4.	4
3	Q	dokumen sedia digital. Kait jiplak cantum sumber referensi.	T1 T2 T3. T4 T5 T6 T7 T8.	8

mencari kemiripan dengan dokumen query.

Tabel 2. Data Pola dan Total term

Dari data diatas maka dapat dihitung nilai similaritasnya adalah:

Sim(Docq, Doc1) = (((3+5)+(3+5))/(8+8))*100% = 100Sim(Docq, Doc7) = ((0+0)/(8+4))*100% = 0

Berikut adalah rancangan tampilan hasil dari deteksi plagiarisme berdasarkan kemiripan pola kalimat dengan threshold kemiripan antar dokumen 50% pada gambar

Nama Pengusul		Mufti Ari			
Judul Topik		Deteksi Plagiat Meng			
Upload Topik Per	nelitian	Choose File No file	chosen		
Thresold Kemirip	oan	50	%	Proses	
Dekumen	Dokumen P	embanding		Presentas: Kemiripan	Has:1
Sumber					

UNIVERSITAS AMIKOM Yogyakarta, 10 Februari 2018

Gambar 8. Rancangan hasil deteksi kemiripan dengan kemiripan pola

Hasil yang diperoleh ditampilkan 1 dokumen yang memliki hasil sangat mirip dengan dokumen *query*, dan dokumen lainnya yang tidak mencapai 50% kemiripan tidak akan ditampilkan pada tabel.

Lain halnya ketika kita menguji kemiripan berdasarkan kesamaan *term* yang terkandung pada tiap kalimat :

Sim(Docq, Doc1)=(8+8)/(8+8))*100%=100 Sim(Docq, Doc7)=((4+4)/(8+4))*100%=66.67

maka rancangan tampilan hasil yang diperoleh adalah seperti pada **Gambar 9**.

Nama Pengusul		Mufti Ari				
udul Topik		De:eksi Plagiat Menggunak				
Uplcad Topik Pe	nelitian	Choose File No file chose	en			
Thresold Kemiripan		50	% Proses			
Dokumen Sumber	Dokumen Per	mbanding	Presentasi Kemiripan	Hasil		
Deteksi Plagiat Menggunakan	Aplikasi Pend Menggunakai Analysis (Stu PCR) docx	100	Sangal Mirip			
LSA.Docx	i chej cana			-		

Gambar 9. Rancangan tampilan Hasil deteksi kemiripan dengan kemiripan *term*.

Hasil yang diperoleh menampilkan 2 buah dokumen yang memiliki kemiripan dengan dokumen *query* dengan presentasi kemiripan 100% sangat mirip untuk dokumen pertama dan 66.67% mirip untuk dokumen kedua.

3. Kesimpulan dan saran

Sistem deteksi plagiarisme topik penelitian menggunakan K-means clustering dan Model Bayesian yang dirancang dapat digunakan untuk menemukan persentasi kemiripan antara dokumen topik penelitian yang diajukan dengan dokumen penelitian sebelumnya. Dari hasil klasifikasi diperoleh dua dokumen yang mendekati dokumen query. Jika hasil kemiripan diuji berdasarkan kesamaan pola kata dalam kalimat, dokumen dengan Nilai similarity tertinggi diperoleh 1 dokumen dengan presentase kemiripan sebanyak 100%, dan jika berdasarkan kesamaan term, berpengaruh terhadap hasil dokumen mirip yang dihasilkan, sehingga diperoleh 2 dokumen mirip. Meskipun uji berdasarkan kesamaan term, menghasilkan dokumen mirip yang lebih banyak namun belum cukup akurat menunjukkan adanya plagiasi, karena dalam menentukan plagiasi kesamaan rangkaian kalimat merupakan hal yang penting untuk diperhatikan.

Disarankan perancangan ini dapat diimplementasikan ke dalam sebuah program dengan dokumen uji yang lebih banyak.

Daftar Pustaka

- Kamus Besar Bahasa Indonesia Daring (Dalam Jaringan).2008. http://bahasa.kemdiknas.go.id/kbbi/index.php. Diakses tanggal 11 Desember 2017.
- [2] Na'riful Hasna Ariyani, Sutardi, RahmadRamadhan, Aplikasi Pendeteksi Kemiripan Isi Teks Dokumen Menggunakan Metode Levenshtein Distance, semanTIK, Vol.2, No.1, Jan-Jun 2016, pp 279-286.
- [3] Lukashenko, R., Graudina, V., dan Grundspenkis, J. "Computer-Based Plagiarism Detection Methods and Tools: An Overview", International Conference on Computer Systems and Technologies, CompSysTech, University of Rousse, Bulgaria, hal. IIIA. 18-(1-2), 2007.
- [4] Kusrini, Tuntunan Praktis Membangun Sistem Informasi Akuntansi dengan Visual Basic dan Microsoft SQL Server, Andi, Yogyakarta, 2007.
- [5] Saunders, E.J. Confroting Academic Dishonesty. Journal of Social Work Education.29(2): 224-31, 1993.
- [6] Utorodewo, Felicia, dkk, Bahasa Indonesia: Sebuah Pengantar Penulisan Ilmiah, FEUI, Jakarta, 2007.
- [7] Riani, Amir Hamzah, Erna Kumalasari N, Klarifikasi Dokumen Berita Bahasa Indonesia Menggunakan Metode Naive Bayes Classifer (NBC) dan K-means Clustering, AKPRIND Yogyakarta, Yogyakarta, 2017.
- [8] S.Agustina, D.Yhudo, H.Santoso, N.Marnasusanto, A.Tirtana, F.Khusnu, *Clustering* Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode *K-Means*, Universitas Brawijaya Malang, Malang, 2012.
- [9] Asroni, Ronald Adrian, 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, 76-82, 2015
- [10] Achmad Syarifudin, Nurul Hidayat, Lutfi Fanani, Sistem Pakar Diagnosis Penyakit Pada Tanaman Jagung Menggunakan Metode Naive Bayes Berbasis Android, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, e-ISSN:2548-964X Vol.2,No.7, 2738-2744, 2018.

Biodata Penulis

Mufti Ari Bianto, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi UNIVERSITAS AMIKOM Yogyakarta, lulus tahun 2017. Saat ini menjadi Mahasiswa di Program Pasca Sarjana Magister Teknik Informatika UNIVERSITAS AMIKOM Yogyakarta.

Siti Rahayu, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika UNIVERSITAS MUSLIM INDONESIA Makassar, lulus tahun 2016. Saat ini menjadi Mahasiswa di Program Pasca Sarjana Magister Teknik Informatika UNIVERSITAS AMIKOM Yogyakarta.

UNIVERSITAS AMIKOM Yogyakarta, 10 Februari 2018

Miftahul Huda, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK WICIDA Samarinda, lulus tahun 2015. Saat ini menjadi Mahasiswa di Program Pasca Sarjana Magister Teknik Informatika UNIVERSITAS AMIKOM Yogyakarta.

Kusrini, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika UNIVERSITAS GADJAH MADA Yogyakarta, lulus tahun 2002. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Teknik Informatika UNIVERSITAS GADJAH MADA Yogyakarta, lulus tahun 2006. Memperoleh Gelar Doktor pada Program Doktoral Teknik Informatika UNIVERSITAS GADJAH MADA, Saat ini menjadi Dosen di STMIK AMIKOM Yogyakarta.