Deteksi Plagiarisme Skripsi Mahasiswa dengan Metode Single-link Clustering dan Jaro-Winkler Distance

Hidayat Abdul Rouf[1], Ardhi Wijayanto[2], Abdul Aziz[3]

[1]*Program Studi Informatika, Universitas Sebelas Maret, Jl. Ir Sutami 36-A, Surakarta, 57126*
E-mail: hidayat@stundent.uns.ac.id
[2]*Program Studi Informatika, Universitas Sebelas Maret, Jl. Ir Sutami 36-A, Surakarta, 57126*
E-mail: ardhi.wijayanto@staff.uns.ac.id
[3]*Program Studi Informatika, Universitas Sebelas Maret, Jl. Ir Sutami 36-A, Surakarta, 57126*
E-mail: aaziz@staff.uns.ac.id

Abstract— The rise of plagiarism is one of the negative impacts of the development of information and communication technology. Plagiarism can occur anywhere. One of the examples is a university with the object of plagiarism as a student's final project. So we need a system to detect plagiarism so that it can suppress plagiarism in the college environment. In detecting the similarity of a writing will be faster if the writing has been grouped before compared to each other. Single-link clustering was chosen because it has a simple algorithm and can be implemented without the initial cluster. In plagiarism plagiarism usually changes the sentence structure so that it looks different Jaro-Winkler distance is chosen because it can detect similarities in paragraphs that have been changed in sentence structure because Jaro-Winkler distance has a flexible indexing with theoretical distance so that a word or character is considered the same. The stages in this study include data collection, preprocessing, grouping writing with Single-link clustering, comparing writing with jaro-winkler distance, and testing with precision and recall. After testing, the average value of precision was 84.37% and recall was 84.37% with a level of plagiarism of 99.1%.

***Keywords*—: jaro-winkler distance; plagiarisme; single-link clustering.**

I. PENDAHULUAN

Kemajuan dalam bidang teknologi informasi dan komunikasi berjalan sangat pesat, salah satu perkembangan yang paling signifikan adalah semakin mudahnya pertukaran data, baik itu dokumen, gambar, maupun suara. Bentuk informasi yang sering digunakan dalam pertukaran data adalah dokumen atau tulisan. Selain membawa dampak positif kemajuan perkembangan teknologi dan informasi juga membawa dampak negatif yaitu plagiarism (Wicaksono, 2012).

Plagiarisme adalah penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikan seolah karangan dan pendapat sendiri (KBBI,1997). Plagiat adalah perbuatan dengan sengaja maupun tidak sengaja mengutip sebagian atau menggunakan seluruh karya orang lain yang diklaim sebagai karyanya tanpa mencantumkan sumber secara tepat dan memadai . Beberapa tipe plagiarisme yaitu :

a. Plagiarisme kata demi kata (*word for word plagiarism*), penulis menggunakan kata-kata penulis lain secara persis.
b. Plagiarisme atas sumber (*plagiarism of sources*), penulis menggunakan gagasan orang lain tanpa memberikan pengakuan yang cukup atau tanpa menyebut sumber yang jelas
c. Plagiarisme kepengarangan (*plagiarisme of authorship*), penulis mengakui sebagai pengarang karya tulis orang lain.
d. *Self plagiarism* penulis mempublikasikan satu artikel pada lebih dari satu redaksi publikasi, dan mendaur uang karya tulis/karya ilmiah.

Plagiarisme menurut presentase kata-kata yang diambil atau dijiplak terbagi menjadi 3 kategori (Sastroasmoro, 2007):

a. Plagiarisme ringan <30%
b. Plagiarisme sedang 30% - 70%
c. Plagiarisme berat >70%.

Dalam dunia pendidikan khususnya universitas rentan terjadi praktik plagiarisme salah satunya plagiarisme dalam pembuatan tugas akhir atau skripsi (Arinda, 2015). Berdasarkan buku panduan penulisan skripsi mahasiswa Informatika UNS tahun 2012, skripsi merupakan karya ilmiah yang disusun oleh mahasiswa berdasarkan hasil penelitian laboratorium, penelitian lapangan dan atau kajian suatu teori dengan bimbingan pembimbing, untuk dipertahankan di hadapan penguji sebagai syarat untuk memperoleh gelar sarjana. Oleh karena itu tugas akhir mahasiswa atau skripsi merupakan dokumen yang sangat penting dan harus dijaga keasliannya dari segala bentuk plagiasi.

Dokumen-dokumen dapat dikelompokkan sesuai dengan dokumen identik dengan dokumen yang lain, pengelompokan ini disebut juga dengan *clustering*. *Clustering* adalah upaya untuk mengelompokkan *record*, observasi, atau mengelompokkan ke dalam kelas yang memiliki kesamaan objek (Mala dkk, 2017). Salah satu metode *clustering* yang biasa digunakan adalah

single-link clustering. *Single-link clustering* memiliki algoritma yang sederhana dan fleksibel karena dapat melakukan *clustering* tanpa memerlukan jumlah inisial cluster seperti kebanyakan algoritma *clustering*.

Kesamaan suatu dokumen atau teks dengan dokumen lainnya dapat diketahui dengan menggunakan pendekatan string *metric* yaitu melakukan perbandingan dua string dengan memasukkan dua string ke dalam fungsi matematis tertentu untuk diketahui jarak antara keduanya. Terdapat beberapa algoritma string *metric* diantaranya adalah hamming *distance*, leivenshtein *distance*, needleman-wunsch *distance*, Jaro-Winkler *distance*, dan sebagainya. Diantara metode yang telah disebutkan algoritma Jaro-Winkler *distance* memiliki kecocokan yang baik dalam pencocokan string yang relatif pendek (Kurniawati, 2010).

Jaro mengenalkan sebuah algoritma perbandingan string yang utamanya digunakan untuk membandingkan nama pertama dan terakhir. Algoritma ini memiliki kecepatan, ketepatan, efektivitas dan nilai yang tinggi dalam membandingkan dua buah string yang memiliki *typographical* yang relatif pendek (Winkler, 2006). Algoritma Jaro menganggap sebuah karakter sama dengan karakter pada string pembanding apabila masih masuk ke dalam jarak teoritis (Jaro, 1989).

Pengembangan dari algoritma Jaro berdasarkan Winkler menggunakan nilai panjang prefiks yang sama di awal string dengan nilai maksimal adalah 4 (Winkler, 2006).

Semakin tinggi Jaro-Winkler *distance* untuk dua string maka semakin mirip kedua string tersebut. Jaro-Winkler *distance* memiliki nilai tertinggi 1 yang berarti kedua string yang dibandingkan sama persis dan nilai terendah 0 yang menandakan kedua string yang dibandingkan tidak memiliki kesamaan sama sekali (Kurniawati, 2010).

Algoritma jaro-winkler distance dapat membandingkan dua stirng dengan cepat dan akurat dengan *indexing* yang lebih fleksibel dibandingkan dengan hamming atau leivenshtein *distance* karena algoritma jaro-winkler *distance* memiliki jarak teoretis agar suatu kata atau karakter dianggap sama.

II. METODE PENELITIAN

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa data skunder yaitu data dokumen tugas akhir mahasiswa. Data dikumpulkan dari program studi Informatika fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sebelas Maret.

B. Preprocessing

Preprocessing adalah tahapan dimana dilakukan seleksi data yang akan diubah menjadi lebih terstruktur sesuai dengan kebutuhan. *Text preprocessing* adalah dimana sebuah dokumen atau teks dipersiapkan menjadi data dengan pengolahan menggunakan cara-cara tertentu sebelum diproses (Mustaqhfiri dkk, 2012). Input awal pada tahap ini adalah dokumen utuh. *Text processing* pada penelitian ini terdiri dari beberapa tahap, yaitu proses *standardization* dan *cleaning*, *stopword removal*, dan *tokenizing*.

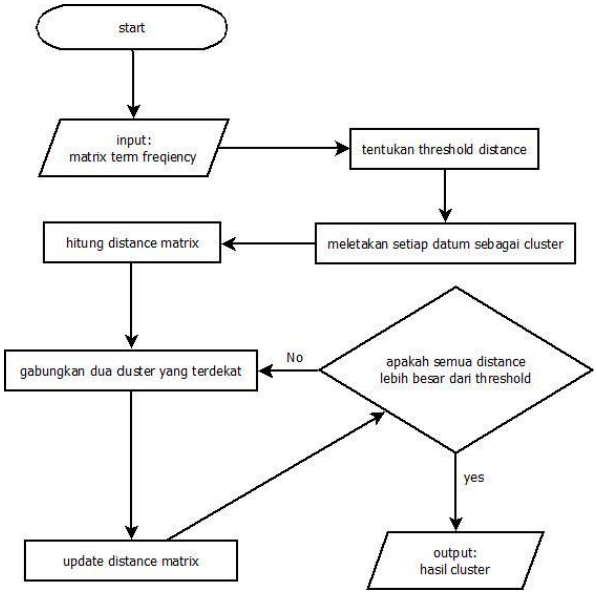
Pada tahap ini dari keseluruhan konten dokumen akan diambil bagian dasar teori kemudian dilakukan *text preprocessing* yaitu *case folding*, *stopword removal* dan *stemming*.

1. *Standardization* dan *Cleaning*
   Standarditation yang digunakan adalah dengan mengubah semua karakter menjadi huruf kecil, dan *cleaning* yang dilakukan adalah dengan menghilangkan semua angka tanda baca dan karakter-karakter yang tidak termasuk dalam ASCII.
2. *Stopword Removal*
   Stopword removal merupakan proses penyaringan kata-kata yang dianggap tidak memiliki arti deskriptif dan tidak memiliki hubungan dengan tema tertentu, kata-kata tersebut sering juga disebut dengan *stopword*.
3. *Tokenizing*
   Tokenization memecah data menjadi per kata atau per simbol. Tahap ini dilakukan dengan menggunakan fungsi RegexpTokenizer yang terdapat pada library nltk di Python.

C. Clustering

Input untuk algoritma Single-link bisa berwujud jarak atau kemiripan antara pasangan-pasangan dari objek-objek. Kelompok-kelompok dibentuk dari entitas individu dengan menggabungkan jarak paling pendek. Metode ini menganggap jarak antara dua cluster sama dengan jarak terpendek antara suatu anggota cluster dengan anggota cluster yang berbeda. Jika data terdiri dari kemiripan, nilai kemiripan antara dua cluster dianggap sama dengan kemiripan terbesar yang dimiliki satu anggota *cluster* dengan anggota *cluster* yang lain (Rokach dan Maimon, 2005).

Tahap clustering menggunakan metode *Agglomerative Clustering* yaitu *Single-link Clustering*. Tujuan dilakukan *clustering* data TA adalah untuk pengelompokan judul *heading* yang berkaitan dan diperkirakan memiliki isi kalimat yang sama.



Gambar 1. Langkah-langkah Proses Single-link Clustering

D. Penentuan Kemiripan

Setelah semua judul *heading* terkelompokkan dilakukan pembandingan atau pencarian kemiripan paragraf dengan paragraf *query* menggunakan Jaro-Winkler *distance*. Paragraf yang dibandingkan adalah paragraf judul *headingnya* satu *cluster* dengan judul *heading query*. Dengan begitu akan mengurangi jumlah proses yang dilakukan namun tidak mengurangi keakuratan karena *string* yang tidak dibandingkan dipercaya tidak memiliki karakteristik yang sama. Langkah menghitung jarak Jaro-Winkler :

1. Menentukan jumlah kata dari *string* 1(s1) dan *string* 2(s2)
2. Menentukan jumlah kata yang dianggap sama dan memenuhi jarak teoritis.

$$\left\lfloor \frac{\max(|s1|,|s2|)}{2} \right\rfloor - 1 \qquad (1)$$

3. Menentukan jumlah transposisi(t), transposisi adalah kata yang ditemukan memiliki kesamaan namun tidak memenuhi jarak teoritis.
4. Menghitung Jaro distance antara s1 dan s2.

$$dj = \frac{1}{3} x \left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{|m|} \right) \qquad (2)$$

5. Menghitung jarak Winkler.

$$dw = dj + (lp(1-dj)) \qquad (3)$$

E. Testing

Pada penelitian ini, dilakukan pengujian untuk mengetahui baik tidaknya metode yang digunakan pada penelitian ini. Pengukuran dilakukan dengan menghitung nilai *recall* dan *precision*.

1. *Recall*

   Recall adalah perbandingan antara jumlah dokumen relevan yang di temukan dengan jumlah dokumen semua dokumen yang relevan di dalam koleksi.

$$recall = \frac{|\{\text{Re}levantDocument\}| \cap |\{\text{Re}trievedDocument\}|}{|\{\text{Re}levantDocument\}|} \qquad (4)$$

2. *Precision*

   Precision adalah perbandingan antara jumlah dokumen relevan yang ditemukan dengan jumlah keseluruhan dokumen yang ditemukan.

$$precision = \frac{|\{\text{Re}levantDocument\}| \cap |\{\text{Re}trievedDocument\}|}{|\{\text{Re}trievedDocument\}|} \qquad (5)$$

III.HASIL DAN PEMBAHASAN

Dari pengumpulan data yang diambil langsung dari jurusan informatika UNS dan dari basis data digilib.uns.ac.id, diperoleh 100 dokumen. Data merupakan dokumen laporan skripsi mahasiswa S1 informatika UNS angkatan 2017-2010. Detail dari data yang diperoleh dapat dilihat pada Tabel 1.

Tabel 1. Detail Data

No	Angkatan	Jumlah
1	2007	28
2	2008	48
3	2009	19
4	2010	5
Total jumlah		100

Sebelum dilakukan *clustering* data yang telah terkumpul akan diproses pada tahap *preprocessing* terlebih dahulu. Hasil dari *text preprocessing* yang terdiri dari penguraian *heading* dan paragraf didapat 788 judul *heading* dan 4924 paragraf.

Proses *clustering* akan mengelompokkan 788 judul *heading* yang telah didapat dari proses pengumpulan data. Hasil dari *clustering* didapatkan 481 *cluster*.

Pengujian dilakukan dengan dokumen laporan skripsi yang telah diurai dengan proses *text preprocessing* dan dihasilkan 4 judul *heading* dan 18 paragraf.

Tabel 2. Detail dokumen query

No.	Nomor Heading	Judul Heading	Jumlah Paragraf
1	2.1.1	Jaro Winler Distance	6
2	2.1.2	Association rule	6
3	2.1.3	Similarity	3
4	2.1.4	Hama dan Penyakit pada Tanaman Padi	3

Judul *heading query* akan dicari di-*cluster* mana dia berada kemudian dibandingkan kemiripan antara paragraf *query* dan paragraf data. Hasil penentuan *cluster* dapat dilihat pada Tabel 2.

Berdasarkan hasil penentuan *cluster*, sistem menentukan hasil yang relevan dan tidak relevan. Relevansi judul heading yang terambil oleh sistem dilihat dari nilai Jaro-Winkler *distance* judul *query* dengan judul *heading* yang terambil maupun judul *heading* di dalam *query*. Hasil perhitungan nilai *precision* dan *recall* terhadap hasil penentuan *cluster* dapat dilihat pada Tabel 4.

Tabel 3. Hasil penentuan cluster

Judul Heading	Cluster	Isi Cluster	id dokumen	id heading
Jaro Winler Distance	110	Jaro Winkler Distance	156	727
		Jaro-Winkler Distance	163	774
Association rule	62	Association rule	156	728
Similarity	17	*Cosine Similarity*	9	68
		Pengukuran *similarity*	107	334
		Similarity	109	361
		Similarity (karhendanan, 2008)	120	441
		Pengukuran *Similarity (*Karhendana, 2008*)*	120	443
		Cosine Similarity	145	642
		Similarity	156	729
		Cosine Similarity	163	779
Hama dan	194	Hama dan	156	730

Penyakit pada Tanaman Padi	Penyakit pada Tanaman Padi

Tabel 4. Hasil precision dan recall

Judul Heading	Retrieved heading	Relevan retrieved heading	Relevan heading dalam library	Preci-sion	recall
Jaro Winler Distance	2	2	2	1	1
Associa-tion rule	1	1	1	1	1
Similarity	8	3	8	0,375	0,375
Hama dan Penyakit pada Tanaman Padi	1	1	1	1	1

Berdasarkan Tabel 4 diperoleh rata-rata *precision* sebesar 84,37% dan rata-rata *recall* sebesar 84,37%. Nilai *presicion* menunjukkan seberapa banyak judul yang relevan dari judul yang berhasil diambil oleh sistem. Sedangkan nilai recall menunjukkan seberapa banyak judul yang relevan dari seluruh data yang ada.

Setelah diperoleh judul *heading* yang satu *cluster* dengan judul *heading query*, sistem kemudian mengambil paragraf yang berada di bawah judul *heading* yang terambil oleh sistem. Paragraf-paragraf ini yang akan dibandingkan dengan paragraf *query* menggunakan persamaan (3). Hasil pembandingan *paragraf query* dengan paragraf library ditunjukkan pada Tabel 5.

Tabel 5. Hasil pembandingan paragraf query dengan paragraf library

Nomor Heading query	Id Paragraf query	Id Paragraf Library	Max Similarity	Match Words	Words Lenght
2.1.1	100	4407	1	49	49
2.1.1	101	4408	1	28	28
2.1.1	102	4409	1	6	6
2.1.1	103	4410	1	124	124
2.1.1	104	4411	0.97	56	60
2.1.1	105	4412	1	53	54
2.1.2	106	4407	1	16	16
2.1.2	107	4408	1	57	57
2.1.2	108	4409	1	170	170
2.1.2	109	4410	1	7	7
2.1.2	110	4411	1	9	9
2.1.2	111	4412	1	6	6
2.1.3	112	2954	1	60	60
2.1.3	113	2955	0.99	87	94
2.1.3	114	2956	1	10	10
2.1.4	115	4407	1	36	36
2.1.4	116	4408	1	36	36
2.1.4	117	4409	1	36	36

Berdasarkan Tabel 5 ditemukan paragraf yang memiliki kemiripan yang tinggi antara paragraf *query* dan paragraf *query* dengan rata-rata nilai similaritas 99.7% dan tingkat plagiarisme dalam dokumen yang diujikan sebesar 99.1%.

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dapat diambil kesimpulan bahwa *Single-link Clustering* dan Jaro-winkler *distance* dapat diimplementasikan pada sistem deteksi plagiarisme dokumen skripsi mahasiswa. Penerapan *Single-link Clustering* digunakan untuk mengelompokkan dokumen sesuai dengan judul *heading* paragraf yang dibandingkan. Kemiripan paragraf dihitung dengan menggunakan Jaro-Winkler *distance*. Untuk mengetahui akurasi dari sistem deteksi plagiarisme ini dilakukan pengujian dengan *precision* dan *recall*. Dari pengujian diperoleh hasil yang baik dengan rata-rata *precision* sebesar 84,37% dan *recall* sebesar 84,37%.

V. UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada kepala program studi S1 Informatika, kedua pembimbing skripsi, kedua orang tua, dan teman-teman yang telah memberikan semangat dan motivasi dalam penyelesaian penelitian ini.

VI. DAFTAR PUSTAKA

Arinda, F.P. (2015). Ketidakjujran Akademik Mahasiswa Perguruan Tinggi X di Surakarta. *J. Appl. Microbiology*., vol. 119, no. 3, 859-867.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, vol. 84, no. 406, 414–420

Kurniawati, A. (2010). Implementasi Algoritma Jaro-Winkler Distance untuk Membandingkan Kesamaan Dokumen Berbahasa Indonesia. In *Seminar Nasional Ilmu Komputer dan Sistem Intelijen KOMMIT 2008*, Depok, Indonesia.

Mala, V, Kusuma, A, Furqon, M. T., dan Muflikhah, L. (2017). Implementasi Metode Fuzzy Subtractive Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan / Lahan. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1, no. 9, 876–884.

Mustaqhfiri, M, Abidin, Z, dan Kusumawati, R. (2012). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. *Matics*, vol. 4, no. 4.

Rokach, L. and Maimon, O. (2005) Clustering Methods. In O. Maimon and L. Rokach (Ed) *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 321–352.

Sastroasmoro, S. (2007). Beberapa Catatan tentang Plagiarisme. *Majelis Kedokt. Indonesia.*, vol. Volum: 57, 239–244.

Wicaksono, Y.A. 2012. *Analisis Dan Implementasi Algoritma Rabin-Karp Dan Algoritma Stemming Nazief-Adriani Pada Sistem Pendeteksi Plagiat Dokumen*.

Winkler, W. E. (2006). Overview of Record Linkage and current research directions. Bureau of The Census., p. 44.