ANALISIS PENGUKURAN *SELF PLAGIARISM* MENGGUNAKAN ALGORITMA RABIN-KARP DAN JARO-WINKLER DISTANCE DENGAN STEMMING TALA

Jayanta[1)], Halim Mahfud[2)], Titin Pramiyati[3)]

[1), 3)] Fakultas Ilmu Komputer UPN "Veteran" Jakarta
[2)]Fakultas Teknik UPN "Veteran" Jakarta
Jl RS. Fatmawati, Pd. Labu, Jakarta 12450
Email : anta.jayanta@gmail.com[1)], halimahfud@upnvj.ac.id[2)], titin.harsono@gmail.com[3)]

Abstrak

Teknologi informasi dan komunikasi (TIK) memiliki kemampuan yang memudahkan penggunanya untuk membuat, menyimpan dan menyebarkan informasi kepada pengguna lain yang membutuhkan. Kapasitas penyimpan data yang besar menjadikan ketersediaan informasi semakin baik. Ketersediaan informasi yang melimpah, tidak hanya memberi dampak yang baik karena kebutuhan informasi dapat diperoleh dengan cepat dan mudah, tetapi memberi dampak buruk dengan adanya praktek plagiarisme. Praktek plagiarisme yang muncul berasal dari lingkungan akademik, sehingga saat ini plagiarisme menjadi permasalahan penting, karena menyangkut hak kekayaan intelektual seseorang. Praktek plagiarisme yang tanpa disadari sering dilakukan adalah mencontek pekerjaan rumah atau tugas ujian yang dilakukan oleh mahasiswa. Deteksi untuk mengetahui adanya plagiarisme pada pekerjaan rumah atau jawaban ujian mahasiswa, dilakukan dengan cara memeriksa kesamaan tulisan pada setiap pekerjaan tersebut. Kesamaan kata dan kalimat pada satu jawaban dengan jawaban lain, akan mengindikasikan adanya plagiarisme. Penggunaan kata atau kalimat yang sama dapat juga dilakukan oleh seorang penulis pada karya tulis yang dibuat, hal ini mungkin terjadi karena adanya materi pembahasan yang saling bersinggungan, atau karya tulis yang dibuat merupakan kelanjutan dari karya tulis sebelumnya. Kata atau kalimat yang pernah dituangkan dalam satu karya ilmiah ini dan digunakan kembali pada karya ilmiah yang lain dianggap sebagai bentuk self-plagiarism. Menentukan adanya self-plagiarism pada sebuah karya ilmiah merupakan tujuan pembahsan topik pada makalah ini. Beberapa aspek yang dianggap dapat mempengaruhi adanya self-plagiarism yang digunakan adalah penggunaan algoritma deteksi kesamaan teks, topik pembahasan yang sama, bagian karya ilmiah yang diuji dalam pencarian tingkat kesamaan teks, dan jumlah kata. Algoritma yang digunakan dalam pengukuran kesamaan teks adalah algoritma Rabin-Karp dan Jaro-Winkler Distance, karya ilmiah yang digunakan sebanyak 3 buah, 2 diantaranya dibuat oleh penulis yang sama dan merupakan topik yang berkelanjutan, menggunakan bagian "pendahuluan" pada karya ilmiah sebagai bagaian yang diuji. Hasil yang diperoleh menunjukkan penggunaan algoritma Rabin-Karp menghasilkan ukuran self-plagiarism yang tinggi pada pengukuran yang menggunakan bagian "pendahuluan". Jumlah kata sangat mempengaruhi hasil pengukuran dengan algoritma Rabin-Karp.

Kata kunci: plagiarisme, *self plagiarism*, *text similarity*, *stemming*, *algoritma Rabin-Karp*, *algoritma Jaro-Winkler Distance.*

1. Pendahuluan

Plagiarisme adalah salah satu bentuk pelanggaran terhadap etika akademis yang disebut sebagai *scientific misconduct* atau *misconduct in science* atau *academic misconduct*. *Scientific misconduct* atau lebih spesifik *research misconduct* diartikan sebagai "fabrikasi, falsifikasi, plagiarisme, atau praktik lain yang sangat menyimpang dari kelaziman dalam komunitas ilmiah dalam pembuatan proposal, pelaksanaan, atau pelaporan penelitian. Hal tersebut tidak termasuk kesalahan yang jujur, seperti ketidaktahuan, kekeliruan mengelola dan menganalisis data, kesalahan interpretasi atau perbedaan pemahaman data [1].

Perbuatan plagiat sangat rawan terjadi di lingkungan pendidikan karena pemberian tugas yang sama pada peserta didik. Pemberian tugas yang sama ini akan menyebabkan terjadinya tindakan copy and paste dalam proses penyelesaian tugas [2]. Kerawanan praktek plagiat di lingkungan pendidikan, mengharuskan pemerintah membuat peraturan tentang plagiat yang dianggap merupakan perbuatan secara sengaja atau tidak sengaja dalam memperoleh nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan/atau karya ilmiah orang lain, tanpa menyatakan sumber secara tepat dan memadai (Mendiknas, 2010).

Sudigdo Sastroasmoro [1] menjelaskan beberapa jenis plagiarisme berdasarkan pada aspek yang diplagiat, pertama, plagiat atas ide, plagiat data penelitian, plagiat kata, plagiat kalimat, plagiat paragraf, dan memplagiat

secara total tanpa melakukan pengubahan apapun. Kedua, plagiat yang dilakukan secara sengaja atau tidak sengaja, seperti memplagiat isi penelitian orang lain. Ketiga, plagiat yang dilakukan pada proporsi/persentasi kata, kalimat, dan paragraf. Sastroasmoro juga menyimpulkan kategori plagiarisme, yaitu kategori ringan memiliki tingkat plagiat sebesar 0–29%, plagiarisme sedang tingkat plagiat sebesa 30–70%, dan plagiarisme berat atau total tingkat plagiat yang dilakukan sebesar 71–100%.

Bentuk plagiarism yang dikenal sebagai self-plagiarism adalah plagiarisme yang dilakukan secara sengaja atau tidak sengaja penggunaaan kembali ide atau kalimat yang pernah dipublikasikan oleh seorang penulis. Self-plagiarism adalah salah satu jenis bentuk plagiarisme dimana penulis mempublikasi ulang keseluruhan atau menggunakan kembali sebagian dari karya ilmiah yang sudah dipublikasi pada karya ilmiah yang baru. *Self-plagirism* kadang dilakukan tanpa disadari oleh penulis dalam penggunaan bahan pustaka pada karya ilmiah yang dibuat.

Kebutuhan untuk mendeteksi adanya praktik *self-plagiarism* pada karya ilmiah, menjadi motivasi dilakukan penelitian pengukuran tingkat kemiripan teks pada makalah karya ilmiah yang dibuat oleh penulis yang sama dengan topik bahasan makalah yang sama juga. Proses deteksi ini dapat dilakukan dengan melakukan pengukuran kesamaan teks atau *text similarity*. Kesamaan (*similarity*) adalah sebuah konsep yang kompleks yang telah banyak dibahas dalam linguistik, filsafat dan teori informasi. Definisi kemiripan setiap bidang memiliki perbedaan, salah satu bidang yang selalu digunakan untuk membuktikan adanya kesamaan tekstual, yaitu dengan mengambil dua atau lebih string dan membandingkan satu sama lain untuk mengetahui adanya kesamaan. Alasan adanya kategori kesamaan teks dikarenakan bahwa manusia berbeda, satu dengan yang lain memiliki ide yang berbeda dan ambang batas yang berbeda, sehingga memungkinkan adanya kesamaan yang dilakukan [4].

Kesamaan teks dapat diketahui melalui proses pengukuran, karena mengukur kesamaan teks memiliki peran penting dalam penelitian yang terkait dengan pembangunan aplikasi sepertu pencarian informasi, deteksi topik, pelacakan topik, klastering dokumen, peringkasan teks dan lainnya. Menemukan kesamaan antara kata adalah bagian mendasar dari kesamaan teks, yang selanjutnya digunakan untuk mendapatkan kesamaan kalimat, paragraph dan kesamaan dokumen [5]. Kesamaan kata dapat terjadi secara leksikal dan semantik, kesamaan kata secara leksikal jika terdapat urutan karakter yang sama. Pengukuran kesamaan kata berbasis string dilakukan berdasarkan pada urutan dan komposisi karakter. Sebuah metrik *string* adalah metrik yang kesamaan dan ketidaksamaan dalam ukuran jarak antara dua *string* teks untuk mendapatkan perkiraan dan perbandingan kecocokan *string* [5].

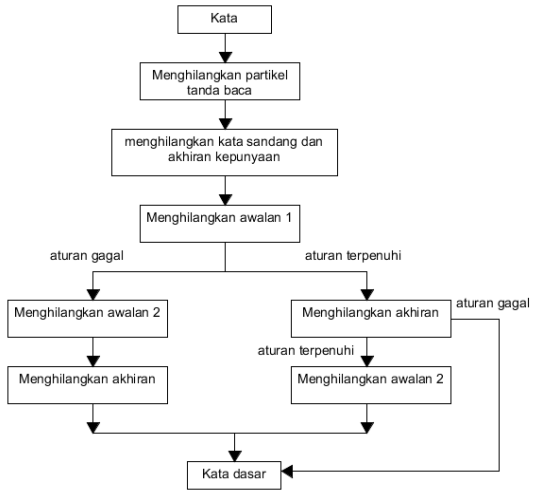
Metode penelitian pengukuran kesamaan teks menggunakan stemming Tala pada algoritma Rabin-Karp dan Jaro-Winkler Distance terdiri dari beberapa tahap, seperti terlihat pada Gambar 1. Tahap pertama adalah pra proses yang terdiri dari proses *Case folding*, yaitu penghilangan huruf kapital pada korpus/dokumen, proses *tokenizing* adalah proses penghilangan tanda baca pada kalimat di dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri, dan proses filtering adalah proses pengambilan kata-kata penting dari hasil *tokenizing*. Algoritma yang digunakan adalah *stop list* (membuang kata yang kurang penting) atau *word list* (menyimpan kata penting).

Gambar 1. *Metode penelitian*

Stemming adalah prosedur komputasi yang mengubah kata menjadi bentuk asalnya (*stem*) dengan mencari awalan, akhiran dan menghapusnya berdasarkan aturan suatu bahasa. Hasil dari proses *stemming* disebut dengan token. Algoritma Tala merupakan algoritma stemming bahasa Indonesia yang dikembangkan oleh Fadillah Z Tala pada tahun 2004[6] dengaan menerapkan *stemming* berbasis aturan atau *rule-based stemmer*. Struktur pembentukan kata dalam Bahasa Indonesia adalah sebagai berikut:

[awalan-1] + [awalan-2] + dasar + [akhiran] + [kepunyaan] + [sandang]

Stemmer Tala merupakan adopsi dari algoritma stemming Porter stemmer. Porter stemmer dipilih berdasarkan pertimbangan bahwa ide dasar stemmer Porter sangat tepat untuk struktur morfologi kata dalam Bahasa Indonesia. Stemming Tala memiliki 5 langkah utama dengan 3 langkah awal dan 2 langkah pilihan, seperti terlihat pada Gambar 2.



Gambar 2. *Langkah proses stemming Tala*

Algoritma Rabin-Karp menggunakan fungsi *hashing* untuk menemukan *pattern* di dalam string teks. Fungsi *hashing* menyediakan metode sederhana untuk menghindari perbandingan jumlah karakter yang quadratik di dalam banyak kasus atau situasi. Pemeriksaan dilakukan dengan cara memeriksa teks yang sedang proses memiliki kemiripan seperti pada *pattern*. Untuk melakukan pengecekan kemiripan antara dua kata ini digunakan fungsi *hash.*

```
function RabinKarp (input s: string[1..m], teks: string[1..n])
boolean
{ Melakukan pencarian string s pada
 string teks dengan algoritma Rabin-K}
i : integer
ketemu = boolean


ketemu ←false
hs ←hash(s*1..m+)
       for i ← 0 to n-m do
         hsub ←hash(teks*1..i+m-1])
         if hsub = hs then
           if teks[i..i+m-1] = s then
             ketemu ←true
           else
             hsub ←hash(teks*i+1..i+m+)
       endfor
return ketemu
```

Algoritma Rabin-Karp ini banyak digunakan dalam pendeteksian pencontekan atau kecurangan. Kelebihan algoritma Rabin-Karp diantaranya sebagai berikut:

1. Rabin-Karp menelusuri karakter satu persatu pada deret karakter, tetapi proses perbandingannya (penghitungan *hash key* nya) relatif mudah.
2. Kasus pencarian string dengan pola yang panjang.

Kekurangan algoritma Rabin-Karp diantaranya sebagai berikut:

1. Membutuhkan waktu yang lama dalam membandingkan kata.
2. Tidak bisa menentukan persamaan makna sinonim kata.

.

Algoritma *Jaro-Winkler distance* yaitu sebuah algoritma untuk mengukur kesamaan antara dua string, biasanya algoritma ini digunakan di dalam pendeteksian duplikat. Semakin tinggi *Jaro-Winkler distance* untuk dua string maka semakin mirip dengan string tersebut. Nilai normalnya ialah 0 menandakan tidak ada kesamaan dan 1 yang menandakan adanya kesamaan. Dasar dari algoritma ini memiliki tiga bagian :

1. Menghitung panjang *string*.
2. Menemukan jumlah karakter yang sama di dalam dua *string*.
3. Menemukan jumlah *transposisi.*

Rumus untuk menghitung jarak (*dj*) antara dua *string* yaitu s_1 dan s_2 seperti pada Persamaan ……(*1*).

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

Keterangan :

m = jumlah karakter yang sama persis.
$|s_1|$ = panjang *string* pertama.
$|s_2|$ = panjang *string* kedua.
t = jumlah *transposisi.*

Alasan penggunaan algoritma Rabin-Karp pada pengukuran ini karena proses pendeteksian kesamaan teks yang digunakan adalah penelusuran karakter pada teks, sedangkan alasan penggunaan algoritma Jaro-Winkler Distance dikarenakan algoritma ini mengukur jarak atau panjang *string* dari dua kata pada proses deteksi kesamaan teks. Perbedaan cara deteksi antara kedua algoritma diduga dapat mempengaruhi hasil pengukuran kesamaan teks sehingga akan mempengaruhi juga prosentase *self-plagiarism* pada dokumen yang diuji.

2. Pembahasan

Penelitian yang dilakukan menggunakan 3 karya ilmiah, yang terdiri dari 2 makalah karya ilmiah ditulis oleh penulis yang sama, dan 1 makalah karya ilmiah milik penulis lain. Semua makalah karya ilmiah membahas topik yang berkaitan dengan pengenalan entitas. Penggunaan 2 makalah karya ilmiah yang ditulis oleh penulis yang sama bertujuan untuk mengukur tingkat *self-plagiarism* yang terjadi pada kedua dokumen.

Makalah ilmiah yang digunakan terdiri dari 3 makalah, yaitu Makala-A, Makalah-B, dan Makalah C sebagai dokumen uji/korpus. Makalah-B dan Makalah-C adalah dokumen yang dibuat oleh penulis yang sama. Makalah-C merupakan makalah lanjutan dari Makalah-B. Ketiga dokumen membahas topik yang sama yaitu topik "pengenalan entitas". Pengukuran tingkat *self-plagiarism* ini menggunakan bagian "pendahuluan" dari makalah dengan alasan pada bagian "pendahuluan" terdapat pembahasan dari keseluruhan isi makalah, dan terdapat

penjelasan teori dari hasil sitasi, sehingga akan terdapat kesamaan teks pada bagian ini. Ukuran dan jumlah kata dari masing-masing makalah sebelum dilakukan pra proses dapat dilihat pada Tabel 1.

Tabel 1. *Makalah karya ilmiah yang digunakan*

No	Nama File	Ukuran File	Bagian Uji Kesamaan	Jumlah Kata
1	Makalah-A.docx	14.67 KB	Pendahuluan	334
2	Makalah-B.docx	29.81 KB	Pendahuluan	1351
3	Makalah-C.docx	29.2 KB	Pendahuluan	839

Hasil pra proses menyisakan 199 kata untuk Makalah-A, 739 kata untuk makalah-B dan 531 kata untuk makalah-C. Pengurangan jumlah kata disebabkan karena adanya proses filtering, yaitu mengambil kata penting saja. Pembuangan kata tidak penting dilakukan menggunakan *stopword*, yaitu kata-kata yang memiliki tingkat kemunculan yang tinggi pada dokumen. Pembuangan kata ini dilakukan dengan cara mencari kata yang sering muncul dari kombinasi dua dokumen, sehingga digunakan skema pembuangan kata A-B, A-C, dan B-C. Penggunaan skema ini menghasilkan jumlah *stopword* yang berbeda untuk masing-masing skema. Skema A-B menggunakan 135 kata, 100 kata untuk skema A-C, dan 127 kata untuk skema B-C.

Hasil proses *stemming* dengan menggunakan *stemmer* Tala menunjukkan terdapat beberapa kata yang mengalami kehilangan suku kata atau huruf pada kata dasar seperti kata "percaya" menjadi "rcaya", hal ini dikarenakan suku kata "pe" dianggap sebagai sebuah awalan. Hal serupa juga terjadi pada kata dengan huruf terakhir "i" seperti kata "notasi", "fungsi" mengalami perubahan karena huruf "i" diakhir kata hilang karena dianggap sebagai akhiran "i". Hasil *stemming* terhadap makalah-1 seperti terlihat pada Gambar 2. Perubahan kata yang terjadi sebagaimana terlihat pada Gambar 2, tidak mempengaruhi proses uji kemiripan teks karena perubahan tersebut terjadi pada kedua makalah.

user profile tampil visual data pribad kaitk guna anggap representas model guna user profile kacu pada representasi digital identitas seseorang eksplisit user profile salah layan sedia layan internet media sosial user profile sedia informas kait jati guna rnyata kenal tinggal riwayat didik fasilitas bagi informas pribad teman rabat gawai dunia guna internet sumber informas potensial manfaat sedia informas asal organisas resmi guna wakil masyarakat partisipas sedia informas kualitas rcaya informas rcaya oleh dasar rcaya milik sumber informas reputas sumber informas rcaya entit mperhati tingkat rcaya trust level milik entit model rcaya bangun tentu tingkat rcaya model rcaya tilai rcaya guna aplikas tentu rcaya guna internet kanisme tentu rcaya reputas

Gambar 2. *Penggalan hasil stemming makalah-1*

Setelah diperoleh kata dasar, tahap berikutnya yang dilakukan adalah uji kemiripan teks dengan menggunakan algoritma Rabin-Karp. Skema peengukuran kesamaan teks menggunakan skema yang sama seperti pada pra proses, yaitu A-B, A-C, dan B-C. Hasil pengukuran seperti terlihat pada Tabel 2. Skema A-B dan A-C menunjukan prosentase kesamaan teks kurang dari 30%, sedangkan kesamaan teks pada skema B-C kmencapai 70%. Besarnya prosentase kesamaan pada skema B-C dipengaruhi oleh kesamaan topik yang ditulis oleh penulis yang sama, jumlah kata yang tidak terlalu besar perbedaannya dan algoritma yang digunakan menggunakan pattern sebagai teknik pendeteksian kesamaan teks.

Tabel 2. *Hasil pengukuran Rabin-Karp*

No	D-1	D-2	Kgram 1	Kgram 2	Kgram sama	Sim (%)	Waktu Proses (s)
1	A	B	707	1929	383	29.06	25.58
2	A	C	707	1468	307	28.23	19.56
3	B	C	1929	1468	1194	70.3	49.85

Hasil pengukuran kesamaan teks dengan algoritma Jaro-Winkler Distance, dapat dilihat pada Tabel 3. Menunjukan hasil pengukuran pada skema B-C sebesar 34.5 (DW). Perbedaan yang cukup besar dengan hasil pengukuran Rabin-Karp dikarenakan penggunaan deteksi kesamaan pada *Jaro-Winkler Distance* berdasarkan pada jarak antar teks. Teks yang sama (s1) pada Makalah-B dan Makalah-C memiliki jarak sama dengan kata lainnya (s2), yang dianggap memiliki kesamaan teks.

Tabel 2. *Hasil pengukuran Jaro-Winkler Distance*

No	D-1	D-2	s1	s2	m	tr	DJ (%)	DW (%)
1	A	B	199	199	48	24	26.8	26.8
2	A	C	199	199	40	20	25.9	25.9
3	B	C	763	763	168	82	34.9	34.5

Daftar Pustaka

[1] S. Sastroasmoro, "Beberapa Catatan tentang Plagiarisme *," pp. 239–244, 2007.
[2] Herqunanto, "Plagiarisme , Runtuhnya Tembok Kejujuran Akademik," *eJKI*, vol. 1, no. 1, pp. 1–3, 2013.
[3] M. P. Nasional, "permendiknas-no-17-tahun-2010_pencegahan-plagiat.pdf." 2010.
[4] A. Ali, "Textual Similarity," Technical University of Denmark, 2011.
[5] W. H. Gomaa, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.
[6] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia."

Biodata Penulis

Jayanta, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi GUNADARMA Jakarta, lulus tahun 1994. Memperoleh gelar Magister Science Komputer (M.Si) Program Pasca Sarjana Magister IPB,

ISSN : 2302-3805

lulus tahun 2007. Saat ini menjadi Dosen di UPN "Veteran" Jakarta.

Titin Pramiyati, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi UPN "Veteran" Jakarta, lulus tahun 1999. Memperoleh gelar Magister Science Komputer (M.Si) Program Pasca Sarjana Magister IPB, lulus tahun 2008. Saat ini menjadi Dosen di UPN "Veteran" Jakarta.

ISSN : 2302-3805