Tersedia di https://jtsiskom.undip.ac.id (31 Januari 2018)
DOI: 10.14710/jtsiskom.6.1.2018.7-12

Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode Jaro-Winkler Distance dan Metode Latent Semantic Analysis

Tinaliah[*,1)], Triana Elizabeth [2)]

[1)]*Program Studi Manajemen Informatika, AMIK MDP*
Jl. Rajawali No 14, Palembang, Indonesia 30113

[2)]*Program Studi Sistem Informasi, STMIK Global Informatika MDP*
Jl. Rajawali No 14, Palembang, Indonesia 30113

Cara sitasi: T. Tinaliah, and T. Elizabeth, "Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode Jaro-Winkler Distance dan Metode Latent Semantic Analysis," Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 1, Jan. 2018. doi: 10.14710/jtsiskom.6.1.2018.7-12, [Online].

Abstract – Various methods are applied in the application of plagiarism detection to help check the similarity of a document. Jaro-Winkler Distance can measure the distance between two strings. However, this method basically depends on the position of the word. Latent Semantic Analysis emphasizes the words contained in the document regardless of its linguistic character. This study compares the results of plagiarism detection using the Jaro-Winkler Distance and the Latent Semantic Analysis method. From comparing results of Jaro-Winkler Distance method and Latent Semantic Analysis method, Jaro-Winkler Distance method is better than Latent Semantic Analysis method if using the same test data. Jaro-Winkler Distance method will give plagiarism result 100% and Latent Semantic Analysis method will give plagiarism result 97,14%.
Keywords - *plagiarism; undergraduate thesis; Jaro-Winkler Distance; Latent Semantic Analysis*

Abstrak – Beragam metode diterapkan dalam aplikasi deteksi plagiarisme untuk membantu mengecek tingkat kesamaan sebuah dokumen. Metode Jaro-Winkler Distance dapat mengukur kesamaan antara dua buah string dan sangat bergantung pada urutan atau posisi kata. Latent Semantic Analysis mementingkan kata-kata yang terkandung di dalam dokumen tanpa memperhatikan karakter linguistiknya. Penelitian ini melakukan perbandingan hasil deteksi plagiarisme dengan menggunakan metode Jaro-Winkler Distance dan metode Latent Semantic Analysis. Hasil pendeteksian plagiarisme dokumen menggunakan metode Jaro-Winkler Distance memberikan hasil yang lebih baik daripada metode Latent Semantic Analysis, yaitu jika data yang dibandingkan sama persis maka akan menghasilkan nilai plagiat sebesar 100%, sedangkan metode Latent Semantic Analysis menghasilkan nilai plagiat sebesar 97,14%.
Kata Kunci - *plagiarisme; karya akhir; Jaro-Winkler Distance; Latent Semantic Analysis*

*) Penulis korespondensi (Tinaliah)
Email: tinaliah@mdp.ac.id

I. PENDAHULUAN

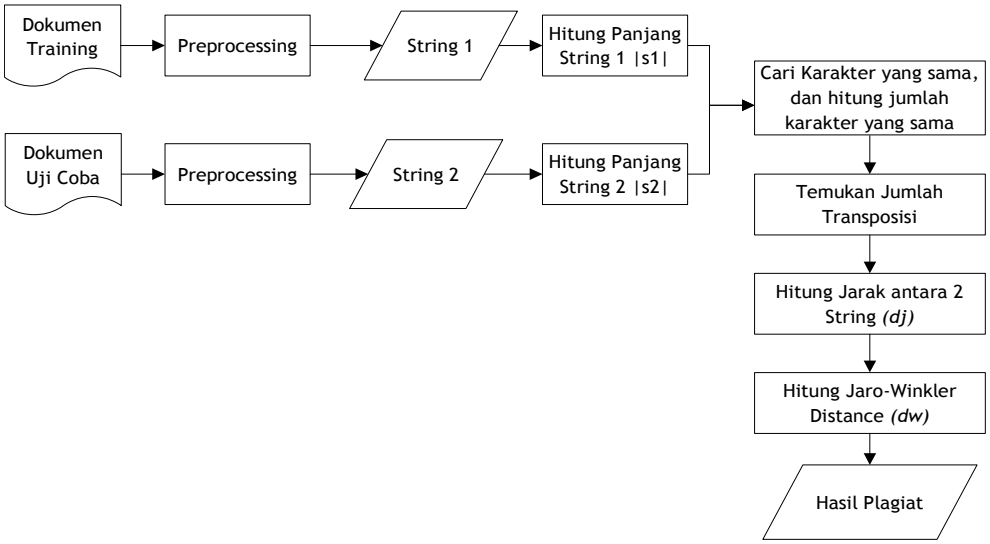
Informasi merupakan sesuatu hal yang sangat penting saat ini, dimana perkembangan teknologi yang semakin pesat menyebabkan informasi semakin terus bertambah. Banyak cara yang dapat dilakukan untuk dapat memperoleh informasi, misalnya melalui buku, majalah, Internet dan sumber-sumber informasi lainnya. Informasi yang diperoleh diharapkan merupakan informasi yang terbaru dan dapat dipercaya. Namun, beberapa orang memanfaatkan informasi berupa karya orang lain untuk diakui sebagai karya ciptaannya sendiri atau melakukan plagiarisme. Plagiarisme dapat dikelompokkan berdasarkan proporsi atau persentase kata, kalimat, atau paragraf yang dibajak, yaitu plagiarisme ringan (<30%), plagiarisme sedang (30 – 70%) dan plagiarisme besar atau total (>70%) [1].

Umumnya praktik plagiarisme terjadi akibat para penulis termasuk mahasiswa terbiasa untuk mengambil bahan tulisan tanpa mencantumkan sumber bahan tersebut berasal. Selain itu, ada juga yang telah mencantumkan sumbernya namun menyalin sama persis dengan sumber, sehingga masih terindikasi sebagai sebuah plagiat. Zulkarnain [2] menyatakan bahwa cara untuk menghindari plagiarisme adalah senantiasa taat pada gaya selingkung, melakukan pengutipan (menyitir) secara langsung, dan melakukan parafrasa terhadap kutipan yang dirujuk.

Contoh praktik nyata plagiarisme di kampus terjadi saat pembuatan skripsi / tesis / disertasi. Mahasiswa yang sedang membuat laporan skripsi, tesis atau disertasi sering mengacu pada bahan-bahan skripsi/tesis/disertasi yang telah dibuat oleh kakak-kakak kelasnya terdahulu. Mahasiswa terbiasa untuk mengambil landasan teori yang ada pada bab 2, sehingga apabila ditelusuri, sebagian besar isi dari landasan teori mereka sama persis. Untuk menghindari plagiarisme yang semakin banyak terjadi, maka diperlukan sebuah aplikasi yang dapat mengecek seberapa besar tingkat plagiarisme sebuah dokumen, agar dapat mengurangi tingkat plagiarisme terutama pada saat pembuatan skripsi / tesis / disertasi.

Metode yang dapat melakukan pendeteksian tingkat plagiarisme sebuah dokumen ada berbagai macam, salah

Dikirim: 27 November 2017; Direvisi: 9 Desember 2017, 25 Januari 2018; Diterima: 26 Januari 2018



Gambar 1. Proses dalam algoritme *Jaro-Winkler Distance* [3]

satunya adalah metode *Jaro-Winkler Distance. Jaro-Winkler Distance* memiliki kelebihan yaitu dapat mengukur kesamaan antara dua *string* dimana metode ini biasanya digunakan di dalam pendeteksian duplikat dokumen. Semakin tinggi nilai *Jaro-Winkler Distance* untuk dua buah *string,* maka semakin mirip *string* tersebut. Nilai normalnya adalah 0 yang menandakan tidak adanya kesamaan dan 1 yang menandakan adanya kesamaan. Metode ini cocok untuk pendeteksian plagiarisme, namun sangat bergantung pada urutan ataupun posisi kata. Metode ini telah digunakan dalam aplikasi di [3]-[5]. Kornain dkk. [3], Kurniawati dkk. [4] dan Faranika dkk. [5] mengembangkan aplikasi pengukur kemiripan dokumen berbahasa Indonesia menggunakan metode *Jaro-Winkler Distance*.

Metode yang lain adalah *Latent Semantic Analysis* (LSA). Metode ini digunakan pada himpunan dokumen yang banyak dan terstruktur untuk mengekstraksi dan mewakili penggunaan arti kata dengan perhitungan statistik dan aljabar linier. LSA merupakan metode yang mementingkan kata-kata yang terkandung di dalam dokumen tanpa memperhatikan karakter linguistiknya. LSA mampu mendeteksi adanya kemiripan makna kata yang diberikan, walaupun kata tersebut muncul atau tidak muncul pada korpus. LSA merupakan suatu metode pembuatan representasi istilah (*term*) berbasis vektor yang dianggap mampu menangkap inti sari dari suatu dokumen maupun kalimat. Metode ini telah digunakan dalam aplikasi di [6]-[8] untuk mengukur kemiripan dokumen berbahasa Indonesia menggunakan metode LSA. Perkasa dkk. [6] menggunakannya untuk sistem ujian online essay, sedangkan Wicaksono dkk. [7], dan Khairunnisa dkk. [8] untuk deteksi kemiripan dokumen akademik. Tudesman dkk. [9] menggunakan metode *Vector Space Model* (VSM) untuk mendeteksi kemiripan dokumen akademik.

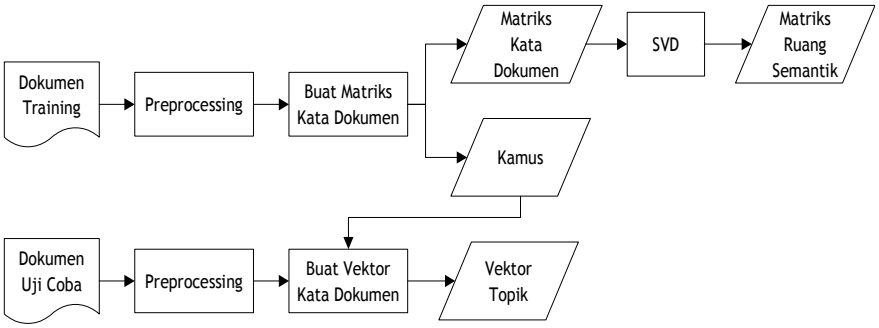
Perbandingan metode deteksi telah dilakukan untuk mengetahui performansi dari metode dalam menganalisis tingkat kesamaan sebuah dokumen, seperti yang dilakukan oleh [10]-[12]. Soleman dan Purwarianti [10] membandingkan antara metode VSM dan LSA. Tinaliah [11] membandingkan deteksi otomatis menggunakan metode LSA dan *centroid-based summarization*. Leonardo dan Hansun [12] melakukan deteksi otomatis menggunakan metode *Rabin-Karp* dan *Jaro-Winkler Distance* berdasarkan efektivitas dan kecepatan prosesnya. Penelitian ini bertujuan menganalisis hasil deteksi tingkat plagiarisme dengan menggunakan metode *Jaro-Winkler Distance* dan metode *Latent Semantic Analysis* untuk menganalisis tingkat kesamaan sebuah dokumen jurnal skripsi mahasiswa dengan dokumen lainnya. Analisis dilakukan terhadap deteksi kesamaan yang melalui proses *stemming* dan tanpa proses *stemming*.

II. METODE PENELITIAN

Data yang digunakan pada penelitian ini terbagi menjadi 2, yaitu :

1. Data latih berjumlah 100 buah dokumen yang digunakan untuk proses pelatihan data pada metode LSA.
2. Data uji coba berjumlah 5 buah dokumen, yaitu :
   a. Abstrak yang diambil sama persis pada salah satu dokumen korpus.
   b. Abstrak yang telah dilakukan perubahan isi.
   c. Abstrak yang berisi paragraf pertama saja pada salah satu dokumen korpus.
   d. Abstrak yang digabungkan dengan isi abstrak lain pada dokumen data latih.
   e. Abstrak yang berbeda dengan dokumen pada korpus.

Data tersebut merupakan data seperti yang digunakan oleh Kornain dkk. [3] dan Tudesman dkk. [9] yang melakukan pengukuran kemiripan dokumen tersebut berurutan dengan menggunakan metode *Jaro-Winkler Distance* dan VSM. Data tersebut adalah berupa dokumen abstrak dari jurnal skripsi Program Teknik Informatika STMIK Global Informatika MDP. Jurnal yang diambil dikonversi ke format *.txt.* Jurnal



Gambar 2. Proses dalam metode LSA

Skripsi diambil dari situs jurnal skripsi mahasiswa/i MDP, yaitu http://eprints.mdp.ac.id/.

Dalam tahap persiapan dokumen, yang dilakukan adalah meliputi tahap *case folding*, tokenisasi, pembuangan *stopwords*, dan *stemming*. Proses dari metode *Jaro-Winkler Distance* ditunjukkan dalam Gambar 1. Metode ini terdiri dari 3 tahap, yaitu menghitung panjang *string,* menemukan jumlah karakter yang sama di dalam dua buah *string* dan menemukan jumlah transposisi [13]. Penghitungan jarak antara dua *string* (d_j)*,* yaitu s_1 dan s_2 dilakukan dengan menggunakan Persamaan 1. Jarak teoretis dua buah karakter yang sama dapat dibenarkan jika tidak melebihi $(\frac{max(|s_1|,|s_2|)}{2})-1$.

$$d_j=\frac{1}{3}x\left(\frac{m}{|S_1|}+\frac{m}{|S_2|}+\frac{m-t}{m}\right) \quad (1)$$

dimana :

m = jumlah karakter yang sama persis
$|s_1|$ = panjang *string* 1
$|s_2|$ = panjang *string* 2
t = jumlah transposisi

Jarak *Jaro-Winkler* (d_w) menggunakan skala prefiks (p) yang menyediakan sebuah prefiks dalam sebuah himpunan *string*. Jarak d_w dinyatakan dalam Persamaan 2. Panjang prefiks (l) menyatakan panjang awalan, yaitu panjang karakter yang sama dengan *string* yang dibandingkan sampai ditemukannya ketidaksamaan (maksimal 4). Jarak *Jaro* untuk *string* s_1 dan s_2 dinyatakan sebagai d_j. Konstanta faktor skala (p) mempunyai nilai standar menurut *Winkler* sebesar *p=0.1*.

$$d_w=d_j+(lp(1-d_w)) \quad (2)$$

Proses pada metode LSA terdiri dari dua tahapan, yaitu tahap pelatihan data dan tahap pendeteksian plagiat. Proses pada metode LSA ditunjukkan pada Gambar 2. Proses metode LSA adalah data latih akan dibuat menjadi matriks kata dokumen, sehingga menghasilkan matriks kata dokumen dan kamus. Matriks kata dokumen akan dilakukan reduksi dimensinya oleh *Singular Value Decompotion* (SVD), sehingga menghasilkan matriks ruang semantik, yang di dalamnya berisi vektor-vektor kata. Data dari dokumen uji coba akan dibuat menjadi vektor kata dokumen sehingga menghasilkan matriks topik sesuai dengan kamus dari data latih. Matriks topik berisi vektor topik yang mewakili topik inti dokumen. Tahap pelatihan dan tahap pendeteksian plagiat dilakukan dengan menggunakan *tools* TMG 5.0R6 yang didapat dari situs *online* TMG, yaitu http://scgroup20.ceid.upatras.gr:8000/tmg/.

Tabel 1. Term yang dijumpai dalam D_1 dan D_2

	Kata	D_1	D_2
k_1	Komputer	1	1
k_2	Kalang	1	1
k_3	Bidang	1	0
k_4	Didik	1	0
k_5	Institusi	1	0

Penilaian yang ada pada metode LSA adalah melakukan perhitungan *cosine similarity* antara matriks ruang semantik dengan vektor topik, sehingga menghasilkan nilai plagiat [14]. *Cosine similarity* digunakan untuk mengukur kedekatan antara dua buah vektor. Persamaan *cosine similarity* ditunjukkan dalam Persamaan 3.

$$\text{sim}(d_j,q)=\frac{d_j.q}{|d_j||q|} \quad (3)$$

dimana :

d_j = dokumen j
q = dokumen *query*

III. HASIL DAN PEMBAHASAN

Cuplikan dokumen yang digunakan sebagai data latih dan data uji coba dalam penelitian ini ditunjukkan dalam Gambar 3 sampai Gambar 8. Secara keseluruhan, dokumen data latih adalah sejumlah 100 buah, sedangkan dokumen uji coba untuk menganalisis kinerja metode *Jaro-Winkler Distance* dan LSA adalah sejumlah 5 buah. Dokumen uji coba ini juga digunakan dalam Kornain dkk. [3] dan dalam Tudesman dkk. [9].

Proses perhitungan pada metode *Jaro-Winkler Distance* dan metode LSA untuk membandingkan kesamaan dua dokumen diuraikan dalam bab ini. Misalnya dokumen D_1 dan D_2 berisi *string* berikut:

D_1 : **Komputer** digunakan banyak **kalangan** di berbagai bidang **pendidikan** dan **institusi**.
D_2 : **Komputer** digunakan banyak **kalangan** untuk bekerja.

```
<DOC>
<DOCID>STMIK-090114-1</DOCID>
<TITLE>
Pengenalan Flora Dan Fauna Kepada Anak Menggunakan Visual Architecting Process TM Alfiarini STMIK MDP Palembang alfiarini@stmik-mdp.net
</TITLE>
<ABSTRAK>
komputer terap kalang bidang institusi perintah swasta didik masyarakat rumah tangga kalang anak anak media proses ajar anak anak tulis kaji komputer media ajar anak kenal flora fauna metodologi kembang perangkat lunak pilih visual architecting process tm tahap teliti implementasi tahap kembang aplikasi piranti program visual basic dasar hasil teliti fitur utama fitur peta indonesia isi informasi peta wilayah indonesia mana flora fauna informasi fitur aplikasi fitur fitur banding butuh anak kenal flora fauna gambar tahap architecting process tm architectural requirement penuh standar guna anak dasar teliti tahap architecturing validation tahap architecting process tm alat bantu std state transition diagram cocok karakteristik pindah layer layer kunci architecting process tm flora fauna visual basic
</ABSTRAK>
</DOC>

<DOC>
<DOCID>STMIK-090114-2</DOCID>
<TITLE>
ANALISA PENGEMBANGAN SISTEM PAKAR UNTUK DIAGNOSA PENYAKIT PADA BURUNG WALET BERBASIS APLIKASI BERGERAK Fredy 2007250031 Nofri 2007250021
</TITLE>
<ABSTRAK>
tuju teliti mengdiagnosa sakit burung walet tanda ganggu burung walet dampak buruk ganggu tumbuh burung walet tulis sistem pakar atas sakit burung walet solusi tangan sistem pakar kembang metodologi kembang sistem rup rational unified process tahap fase inception elaboration construction transition aplikasi rancang bahasa program java tentu sakit burung walet sistem pakar proses konsultasi sistem guna harap sistem pakar informasi tangan anak deteksi alami ganggu burung walet kunci analisis sistem pakar rup diagnosa sakit burung walet
</ABSTRAK>
</DOC>
```

Gambar 3. Dokumen data latih

Komputer telah digunakan dan diterapkan oleh banyak kalangan di berbagai bidang, mulai dari institusi pemerintahan, swasta, pendidikan, masyarakat umum, rumah tangga dan sebagainya termasuk untuk kalangan anak-anak sebagai media proses pembelajaran untuk anak-anak. Tulisan ini akan mengkaji bagaimana komputer digunakan sebagai media pembelajaran anak untuk mengenal flora dan fauna. Metodologi pengembangan perangkat lunak yang dipilih adalah Visual Architecting Process TM yang terdiri dari lima tahap. Peneliti mengimplementasi tiap tahapnya dalam pengembangan aplikasi ini dengan piranti untuk pemrograman menggunakan Visual Basic.
Berdasarkan hasil penelitian terdapat dua fitur utama yakni fitur Peta Indonesia yang berisi informasi peta wilayah Indonesia dimana flora dan fauna yang akan diinformasikan berada dan fitur Keluar untuk keluar dari aplikasi.
Fitur-fitur ini kemudian dibandingkan dengan kebutuhan anak terutama pengenalan flora dan fauna yang tergambar pada tahap kedua Architecting Process TM yakni Architectural Requirement apakah telah memenuhi standar yang diinginkan pengguna dalam hal ini anak. Berdasarkan penelitian pada tahap Architecturing validation pada tahap Architecting Process TM bahwa alat Bantu STD (State Transition Diagram) cocok untuk digunakan karena karakteristik dapat memperlihatkan perpindahan dari layer satu ke layer lain.
Kata Kunci: Architecting Process TM, flora, fauna, Visual Basic

Gambar 4. Dokumen uji coba 1

Komputer telah digunakan dan diterapkan oleh banyak kalangan di berbagai bidang, mulai dari institusi pemerintahan, swasta, pendidikan, masyarakat umum, rumah tangga dan sebagainya termasuk untuk kalangan anak-anak sebagai media proses pembelajaran untuk anak-anak. Tulisan ini akan mengkaji bagaimana komputer digunakan sebagai media pembelajaran anak untuk mengenal flora dan fauna. Metodologi pengembangan perangkat lunak yang dipilih adalah Visual Architecting Process TM yang terdiri dari lima tahap. Peneliti mengimplementasi tiap tahapnya dalam pengembangan aplikasi ini dengan piranti untuk pemrograman menggunakan Visual Basic.
Berdasarkan hasil penelitian terdapat dua fitur utama yakni fitur Peta Indonesia yang berisi informasi peta wilayah Indonesia dimana flora dan fauna yang akan diinformasikan berada dan fitur Keluar untuk keluar dari aplikasi.
Kata Kunci: Architecting Process TM, flora, fauna, Visual Basic.

Gambar 5. Dokumen uji coba 2

Komputer telah digunakan dan diterapkan oleh banyak kalangan di berbagai bidang, mulai dari institusi pemerintahan, swasta, pendidikan, masyarakat umum, rumah tangga dan sebagainya termasuk untuk kalangan anak-anak sebagai media proses pembelajaran untuk anak-anak. Tulisan ini akan mengkaji bagaimana komputer digunakan sebagai media pembelajaran anak untuk mengenal flora dan fauna. Metodologi pengembangan perangkat lunak yang dipilih adalah Visual Architecting Process TM yang terdiri dari lima tahap. Peneliti mengimplementasi tiap tahapnya dalam pengembangan aplikasi ini dengan piranti untuk pemrograman menggunakan Visual Basic.

Gambar 6. Dokumen uji coba 3

Komputer telah digunakan dan diterapkan oleh banyak kalangan di berbagai bidang, mulai dari institusi pemerintahan, swasta, pendidikan, masyarakat umum, rumah tangga dan sebagainya termasuk untuk kalangan anak-anak sebagai media proses pembelajaran untuk anak-anak. Chatting merupakan salah satu cara komunikasi yang dilakukan dua orang atau lebih.
Pada aplikasi chatting berbasis teks antar ponsel ini, penulis memanfaatkan teknologi bluetooth yang merupakan media koneksi wireless yang tersedia pada ponsel sebagai media pengiriman data. Metode yang digunakan dalam penelitian ini adalah metode korelasi sederhana dengan menggunakan bantuan program aplikasi SPSS (Statistical Product and Service Solutions). Aplikasi ini dibangun dengan memperincikan berbagai aspek keamanan data yang tersimpan di dalamnya sehingga pengguna aplikasi ini harus didaftarkan terlebih dahulu oleh administrator agar tidak sembarang orang yang bisa menggunakan aplikasi ini.
Kata kunci : Analisis, sistem pakar, RUP, diagnosa penyakit pada burung walet

Gambar 7. Dokumen uji coba 4

Masa remaja merupakan masa transisi antara masa kanak-kanak menuju dewasa, sedangkan keluarga merupakan unit sosial terkecil yang memberikan fondasi primer bagi perkembangan anak, maka dari itu keutuhan keluarga memiliki peranan penting dalam mempengaruhi perkembangan anak atau remaja. Namun tidak demikian pada anak korban perceraian. Berbagai dampak psikis dan sosial dapat timbul sebagai akibat dari perceraian orangtua. Salah satunya adalah hambatan yang dialami remaja saat berkomunikasi interpersonal. Komunikasi interpersonal sendiri adalah komunikasi mendalam yang terjadi secara langsung dengan oranglain dan saling mempengaruhi satu sama lain. Penelitian ini bertujuan untuk mengetahui hambatan komunikasi interpersonal pada remaja korban perceraian. Pada penelitian ini digunakan penelitian kualitatif berbentuk studi kasus karena peneliti ingin mendeskripsikan lebih mendalam mengenai komunikasi interpersonal pada remaja korban perceraian. Metode yang digunakan dalam penelitian ini adalah metode wawancara dan field note yaitu peneliti mencatat dengan menguraikan tentang apa yang terjadi di lapangan, sesuai dengan fokus penelitian, ditulis secara deskriptif dan reflektif. Subjek yang digunakan adalah seorang remaja perempuan berusia 21 tahun yang telah mengalami perceraian orangtua. Subjek dalam penelitian ini memiliki keterbatasan dalam berkomunikasi interpersonal. Tercermin dari sikapnya yang cenderung pendiam dan hanya terbuka pada satu orang saja. Sedangkan hal-hal yang mempengaruhi komunikasi subjek menjadikan subjek rendah diri, pengalaman yang buruk dengan teman-temannya dan sikap defensif sehingga subjek tidak memiliki teman untuk berbagi kecuali dengan kekasihnya.
Kata Kunci : Komunikasi Interpersonal, Remaja, Perceraian.

Gambar 8. Dokumen uji coba 5

Dokumen pada D_1 dan D_2 akan dilakukan *preprocessing* melalui tahapan *case folding*, tokenisasi, proses *stemming* dan pembuangan *stopword.* Hasil *preprocessing*-nya adalah sebagai berikut:

D_1 : **komputer kalang bidang didik institusi**.

D_2 : **komputer kalang**

Dalam perhitungan pada metode *Jaro-Winkler Distance,* dokumen D_1 dianggap sebagai s_1, dan dokumen D_2 dianggap sebagai s_2. Untuk kasus dalam dokumen D1 dan D2 tersebut di atas, maka nilai parameternya adalah sebagai berikut:

$m = 2$

$|s_1| = 5$

Tabel 2. Hasil pengujian melalui proses *stemming*

Pengujian	*Jaro-Winkler Distance*	LSA
Uji Coba 1	100%	97,14%
Uji Coba 2	92,29 %	93,32%
Uji Coba 3	88,32%	86,91%
Uji Coba 4	69,28%	26,43%
Uji Coba 5	28,95 %	26%

Tabel 3. Hasil pengujian tanpa melalui proses *stemming*

Pengujian	*Jaro-Winkler Distance*	LSA
Uji Coba 6	100%	97,53%
Uji Coba 7	92,29%	94,64%
Uji Coba 8	88,32 %	87,16%
Uji Coba 9	70,04%	35,64%
Uji Coba 10	32.5 %	33,22%

$|s_2| = 2$
$t = 0$

Nilai jarak *Jaro*-nya, d_j , dengan menggunakan Persamaan 1 adalah:

$$d_j = \frac{1}{3} x \left(\frac{2}{5} + \frac{2}{2} + \frac{2-0}{2} \right) = 0{,}792$$

Jika diperhatikan susunan s_1 dan s_2 dapat diketahui nilai l = 2, dan nilai konstan p = 0,1. Nilai jarak *Jaro-Winkler,* d_w, dengan menggunakan Persamaan 2 adalah:

$$d_w = 0{,}792 + (2 x 0.1 (1 - 0{,}792)) = 0{,}833$$

Jadi, dokumen D_1 dan D_2 mempunyai kemiripan sebesar 0,833 atau 83,3 %.

Hasil perhitungan pada metode LSA menggunakan *cosine similarity* dinyatakan dalam Tabel 1. Vektor D_1 bernilai $1k_1 + 1k_2 + 1k_3 + 1k_4 + 1k_5$, sedangkan vektor D_2 bernilai $1k_1 + 1k_2 + 0k_3 + 0k_4 + 0k_5$. Perhitungan *cosine similarity* menggunakan Persamaan 3 adalah sebagai berikut:

$$\text{sim}(D_1, D_2) = \frac{(1x1)+(1x1)+(1x0)+(1x0)+(1x0)}{\sqrt{1^2+1^2+1^2+1^2+1^2} x \sqrt{1^2+1^2+0^2+0^2+0^2}} = 0{,}632$$

Jadi, dokumen D_1 dan D_2 mempunyai kemiripan sebesar 0,632 atau 63,2 %.

Pengujian yang dilakukan pada penelitian ini adalah membandingkan kelima dokumen uji coba dengan dokumen pertama data latih menggunakan metode *Jaro-Winkler Distance* dan metode LSA. Pengujian akan dilakukan sebanyak 10 kali, dengan menggunakan 5 data uji coba yang akan dibandingkan dengan dokumen yang isinya sama persis dengan data uji coba pertama. Proses pengujian akan menilai hasil deteksi plagiarisme melalui proses *stemming* dan tanpa melalui proses *stemming*.

Hasil pengujian yang telah dilakukan ditunjukkan pada Tabel 2 dan Tabel 3. Metode *Jaro-Winkler Distance,* dengan algoritme sesuai [13], dapat digunakan untuk mendeteksi kemiripan teks berbahasa Indonesia dan memberikan hasil plagiat sebesar 100%, jika menggunakan data uji coba yang sama dengan data yang dilatih. Secara keseluruhan, metode *Jaro-Winkler Distance* dengan *stemming* mempunyai nilai deteksi yang lebih baik daripada tanpa *stemming*, seperti yang telah dihasilkan dalam [3]. Dalam penelitian ini tidak dilakukan pengukuran waktu yang dibutuhkan untuk melakukan deteksi seperti halnya dalam [5], yang menyatakan bahwa lamanya waktu pengecekan ditentukan oleh ukuran, tipe dan kandungan isi dokumen. Penambahan tahap *stemming* yang memberikan kinerja lebih baik dalam deteksi dapat diberikan ke aplikasi deteksi plagiarisme berbasis web dalam [4].

Metode LSA dalam [14] juga telah dapat mendeteksi kemiripan antar dokumen teks berbahasa Indonesia, seperti halnya dalam [6]-[8]. Wicaksono dkk. [7] telah menambahkan model Bayesian untuk menjaga urutan term / kata sehingga hasil deteksi kemiripan yang dihasilkan dapat lebih baik. Seperti halnya dalam Tudesman dkk. [9] yang menggunakan VSM, penelitian ini juga menganalisis pengaruh pemberian tahap *stemming* dalam proses deteksi menggunakan LSA dimana proses *stemming* dapat memberikan hasil deteksi yang lebih baik.

Dari hasil dari pengujian kedua metode tersebut dapat dinyatakan bahwa pada hasil pengujian metode *Jaro-Winkler Distance* menghasilkan hasil deteksi plagiarisme yang lebih baik daripada metode LSA. Jika dibandingkan dengan [10] yang menyatakan LSA lebih baik daripada VSM, maka dapat dinyatakan juga bahwa *Jaro-Winkler Distance* ini memberikan hasil deteksi kemiripan yang lebih baik daripada VSM. Namun, apabila dokumen telah dimodifikasi, metode LSA memberikan hasil deteksi yang lebih baik sebesar 93,32% pada uji coba 2. Metode LSA ini dapat memberikan hasil deteksi yang lebih baik jika digabungkan dengan *centroid-based summarization* [11].

Penelitian ini telah menghasilkan analisis perbandingan metode *Jaro-Winkler Distance* dan LSA yang menggunakan proses *stemming* dan tanpa *stemming* untuk mendeteksi kemiripan. Perbandingan metode dinyatakan berdasarkan kemampuan dalam deteksi kemiripan. Penelitian lebih lanjut dapat dilakukan untuk membandingkan efektivitas dan kecepatan proses kedua metode seperti yang dilakukan oleh Leonardo dan Hansun [12].

IV. KESIMPULAN

Berdasarkan hasil pengujian pada penelitian ini dapat disimpulkan bahwa hasil *Jaro-Winkler Distance* memberikan hasil yang lebih baik daripada metode *Latent Semantic Analysis*, yaitu jika data yang dibandingkan sama persis maka akan menghasilkan nilai plagiat sebesar 100%, sedangkan metode *Latent Semantic Analysis* menghasilkan nilai plagiat sebesar 97,14%. Hasil pendeteksian plagiarisme dokumen

menggunakan metode *Latent Semantic Analysis* memberikan hasil yang baik apabila dokumen telah dimodifikasi.

DAFTAR PUSTAKA

[1] S. Sastroasmoro, "Beberapa Catatan tentang Plagiarisme," *Majalah Kedokteran Indonesia*, vol. 56, no. 1, pp. 1-6, 2006.

[2] Z. Zulkarnain, "Plagiarisme Dalam Menghasilkan Karya Tulis Ilmiah," April 2013. [Online]. Available: http://www.unja.ac.id/2013/04/10/prof-dr-ir-h-zulkarnain-mhortsc/. [Diakses: Nov, 15, 2017]

[3] A. Kornain, F. Yansen, and T. Tinaliah, "Penerapan Algoritma Jaro-Winkler Distance Untuk Sistem Deteksi Plagiarisme pada Dokumen Teks Berbahasa Indonesia," Skripsi, STMIK MDP, Oktober 2014.

[4] A. Kurniawati, S. Puspitodjati, and S. Rahman, "Implementasi Algoritma Jaro-Winkler Distance untuk Membandingkan Kesamaan Dokumen Berbahasa Indonesia", Skripsi Program Studi Sistem Informasi, Universitas Gunadarma, 2010.

[5] Y. Faranika, H. Kurniawan, and N. Nikentari, "Sistem Pengukur Kemiripan Dokumen Menggunakan Algoritma Jaro-Winkler Distance," Skripsi, Universitas Maritim Raja Ali Haji, 2014.

[6] D. A. Perkasa, E. Saputra, and M. Fronita, "Sistem Ujian Online Essay dengan Penilaian Menggunakan Metode Latent Semantic Analysis (LSA), " *Jurnal Rekayasa dan Manajemen Sistem Informasi*, vol. 1, no. 1, Februari 2015, pp. 1-9.

[7] D. W. Wicaksono, M. I. Irawan, and A. M. Rukmi, "Sistem Deteksi Kemiripan Antar Dokumen Teks Menggunakan Model Bayesian pada Term Latent Semantic Analysis (LSA)," *Jurnal Sains dan Seni POMITS*, vol. 3, no. 2, 2014, pp.41-46.

[8] N. Khairunnisa, D. S. Sihabudin, and A. Wibowo, "Aplikasi Pendeteksi Plagiat dengan Menggunakan Metode Latent Semantic Analysis (Studi Kasus : Laporan TA PCR)," *Jurnal Aksara Komputer Terapan*, vol. 1, no. 2, 2012.

[9] T. Tudesman, E. Oktalina, T. Tinaliah, Y. Yoannita, "Sistem Deteksi Plagiarisme Dokumen Bahasa Indonesia Menggunakan Metode Vector Space Model," Skripsi, STMIK MDP, Oktober 2014.

[10] S. Soleman, and A. Purwirianti, "Experiments on the Indonesian Plagiarism Detection using Latent Semantic Analysis," in *2014 2nd International Conference on Information and Communication Technology (IcoICT)*, 28-30 May 2014, Bandung, Indonesia.

[11] T. Tinaliah, "Ringkasan Multi-Dokumen Berbahasa Indonesia Secara Otomatis Menggunakan Metode Latent Semantic Analysis dan Centroid-Based Summarization," Tesis, Universitas Indonesia, 2013.

[12] B. Leonardo, and S. Hansun, "Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. 2, pp. 462-471, 2017.

[13] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Methods*, American Statistical Association, 1990, pp. 354-359.

[14] S. T. Dumais, "Latent Semantic Analysis," *Annual Review of Information Science and Technology,* vol. 38, no. 1, pp. 188-230, 2004.