PENGUKUR SEMANTIC SIMILARITY PADA ARTIKEL WEB DALAM UPAYA PENCEGAHAN PLAGIARISME

Anacostia Kowanda[1]
Ika Pretty Siregar[2]
Junior Lie[3]
Nur Farida Irmawati[4]
Detty Purnamasari[5]

[1,2,3,4]JurusanTeknik Informatika, Universitas Gunadarma
[5]JurusanSistemInformasi, UniversitasGunadarma
e-mail : ([1]anacos, [5]detty) @staff.gunadarma.ac.id
([2]ikaps, [3]juniorlie, [4]farida_28) @student.gunadarma.ac.id

Abstrak

Pada perkembangan teknologi saat ini, penulisan dokumen secara digital sudah banyak dilakukan. Beberapa instansi atau organisasi pun mengharapkan para anggotanya untuk menulis dokumen dalam bidang tertentu secara digital, salah satunya dalam bentuk artikel. Banyaknya penulis yang sudah membuat beberapa artikel dalam bidang yang berbeda secara digital pada web, menyebabkan terjadinya berbagai kemungkinan bagi para penulis lain untuk melakukan tindakan plagiat dalam penulisan artikelnya. Bagi pihak instansi atau organisasi yang bertugas untuk menilai artikel para penulis membutuhkan sebuah sistem untuk menguji keaslian dari artikel tersebut. Pada paper ini dilakukan penelitian tentang pendeteksian tindakan plagiat pada dua buah artikel web yang berbeda. Penelitian ini dilakukan menggunakan metode Latent Semantic Analysis. Metode tersebutdilakukan dengan menghitung TF-IDF masing-masing term pada masing-masing artikel dan menghitung kemiripan artikelnya dengan Cosine Similarity. Dalam penelitian ini akan diuji tiga artikel dengan satu artikel penguji, dengan menemukan nilai Cosine Similarity tertinggi maka dapat disimpulkan artike ltersebut memiliki tingkat kesamaan terbesar dengan artikel penguji.

Kata kunci : *Latent Semantic Analysis, Vector Space Model,* Artikel, Web

PENDAHULUAN

Kemajuan teknologi yang semakin berkembang saat ini memberikan beberapa perubahan dalam penyajian dokumen, salah satunya penyajian dalam bentuk digital yaitu web. Dengan fasilitas internet yang mudah didapat pada masa sekarang ini memudahkan para penulis untuk mencari sumber-sumber tulisan dalam pembuatan artikelnya.Banyak instansi atau organisasi yang memanfaatkan teknologi sebagai media untuk menyajikan dokumen secara digital pada web. Semakin banyak penulis yang menulis artikel pada web maka semakin banyak pula kesempatan untuk melakukan tindakan plagiarisme. Beberapa tindakan plagiarisme tersebut masih belum terdeteksi oleh beberapa instansi atau organisasi yang menilai tentang keaslian artikel. Biasanya proses pendeteksian

kedua artikel tersebut masih dilakukan secara manual dengan melihat isi dari kedua artikel tersebut lalu dibandingkan. Namun cara tersebut cukup memakan waktu lama dan kemungkinan terjadi kesalahan dalam perbandingannnya sangat mungkin terjadi. Maka dari itu dalam penelitian ini akan dibuat sebuah aplikasi untuk mengukur *Semantic Similarity* pada artikel web dalam upaya pencegahan plagiarisme yang dilakukan oleh para penulis yang kurang bertanggung jawab. Pada artikel ini mengembangkan metode *Latent Semantic Analysis* dengan perhitungan *Cosine Similarity.*

Gambar 1 merupakan rancangan sistem yang dilakukan. Berikut akan dijelaskan masing-masing prosesnya :

a.Masukkan Link

Pada tahap ini, user memasukkan link ke search bar yang tersedia. Terdapat dua search bar, pertama untuk link dengan isi artikel sebagai penguji dan kedua diisi dengan web dimana menyimpan artikel yang diuji.

b.Simpan Artikel dalam Format Teks

Pada langkah ini, sistem akan mengambil artikel atau postingan yang ada pada web kemudian menyimpannya dalam bentuk teks.

Gambar 1. Perancangan Pembuatan Sistem

c.Tokenisasi, Stopwords Removal dan Stemming.

1.Tokenisasi

Tokenisasi merupakan suatu proses pemisahan setiap kata yang terdapat dalam suatu artikel. Di tahap ini akan dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya [4]. Langkah-langkah melakukan tokenisasi :

1. Carilah tanda baca yang ada di dalem artikel (seperti titik, koma,titik dua, titik koma, tanda seru, tanda tanya dan sebagainya).
2. Hilangkan semua tanda baca tersebut.
3. Susun dan rapikan kembali artikel.

2.Stopwords Removal

Tahap ini merupakan proses penghilangan kata-kata tidak penting (stopwords) dari suatu artikel. Proses penghilangan dilakukan dengan pencocokan kata-kata yang ada dalam artikel terhadap sebuah tabel yang berisikan daftar kata-kata tidak penting [5]. Langkah-langkah melakukan stopword removal :

1. Carilah kata umum yang tidak memiliki makna di dalam artikel (pada umumnya merupakan preposisi, seperti yang, di, ke, dari, dan sebagainya).
2. Hilangkan semua stopword tersebut.
3. Susun dan rapikan kembali artikel.

3. Stemming

Stemming merupakan proses pengembalian suatu kata berimbuhan ke bentuk dasarnya (root word). [5]. Langkah-langkah melakukan stopword removal :

1. Carilah kata yang memiliki imbuhan dan akhiran lalu ubahlah menjadi kata dasarnya (contoh : merubah >> ubah, melupakan >> lupa, dan lain-lain).
2. Susun dan rapikan kembali artikel.

d.Perhitungan

Selanjutnya dilakukan perhitungan terhadap artikel-artikel tersebut dengan langkah-langkah sebagai berikut:

1.Tentukan bobot untuk setiap term dari artikel-artikel tersebut dengan menggunakan rumus:

$$idf = \log\left(\frac{n}{df}\right)$$

$$Wdt = tf * idf$$

2.Hitung hasil perkalian scalar artikel yang diuji. Lalu jumlahkan hasil perkalian dari setia partikel diuji dengan artikel penguji.

3. Hitung panjang setiap artikel, dengan cara mengkuadratkan bobotsetiap term yang didapat dari langkah 1 untuk setiap artikel, jumlahkan nilai kuadrat lalu akarkan.

4.Hitung kemiripan vector dari artikel penguji dengan setiap artikel yang diuji dengan menggunakan rumus cosine similarity:

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{ij})}{\sqrt{\sum_k d^2{}_{ik}}\sqrt{\sum_k d^2{}_{jk}}}$$

HASIL DAN PEMBAHASAN

Tabel 1 menyajikan implementasi langkah-langkah Metode Penelitian untuk menguji artikel D1, D2, dan D3 terhadap artikel penguji D0.

Selanjutnya lakukan tokenisasi, stopword removal dan stemming, sehingga didapat hasil seperti pada tabel 2.

tabel yang berisi Term, lalu cari tf dan idf (log n/df) untuk masing-masing term.

Ket : term merupakan kata - kata yang terdapat dalam artikel .

Tabel 1. Langkah Perhitungan 1

D0	Logo Turnamen
D1	Piala dunia tahun ini akan diakan di negara Brazil. Ini adalah kedua kalinya Brazil menjadi tuan rumah turnamen ini.
D2	Tim sepakbola dari 31 negara telah lolos kualifikasi untuk mengikuti piala dunia yang diselenggarakan di Brazil pada bulan Juni 2014.
D3	Brazil yang menjadi tuan rumah piala dunia 2014, telah menyediakan logo untuk turnamen resmi berjudul "Inspiration".

Tabel 2. Langkah Perhitungan 2

	D0	logo turnamen
D1	D1	piala dunia tahun ada negara brazil dua kali brazil jadi tuan rumah turnamen
	D2	timsepakbola 31 negara lolos kualifikasi ikut piala dunia selenggara Brazil bulan Juni 2014
D3	D3	brazil jadi tuan rumah piala dunia 2014 sedia logo turnamen resmi judul Inspiration

Tabel 3. Langkah Perhitungan 3

	tf					idf
Term	D0	D1	D2	D3	df	log(n/df)
logo	1			1	2	0.301
turnamen	1	1		1	3	0.124
piala		1	1	1	3	0.124
dunia		1	1	1	3	0.124
tahun		1			1	0.602
ada		1			1	0.602
Brazil		1	1	1	3	0.124
negara		1	1		2	0.301
jadi		1		1	2	0.301
tuan		1		1	2	0.301
rumah		1		1	2	0.301
tim			1		1	0.602
sepakbola			1		1	0.602
31			1		1	0.602
lolos			1		1	0.602
kualifikasi			1		1	0.602
ikut			1		1	0.602
selenggara			1		1	0.602
bulan			1		1	0.602
Juni			1		1	0.602
2014			1		1	0.602
sedia				1	1	0.602
resmi				1	1	0.602
judul				1	1	0.602
Inspiration				1	1	0.602

Keterangan :
n : 4, yaitu D0, D1, D2, dan D3.
tf : banyaknya kata dalam suatu artikel.
df : total artikel yang mengandung suatu kata.

Selanjutnya hitung nilai bobot (Wdt) dengan mengalikan tf dengan idf

Tabel 4. Langkah Perhitungan 4

wdt=tf.idf			
D0	D1	D2	D3
0.301			0.301
0.124	0.124		0.124
	0.124	0.124	0.124
	0.124	0.124	0.124
	0.602		
	0.602		
	0.124	0.124	0.124
	0.301	0.301	
	0.301		0.301
	0.301		0.301
	0.301		0.301
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
		0.602	
			0.602
			0.602
			0.602
			0.602

Lalu berikutnya, hitung panjang vektor menggunakan Vector Space Model (VSM).

Tabel 5. Langkah Perhitungan 5

WD*Wdi			Panjang Vektor			
D1	D2	D3	D0	D1	D2	D3
0	0	0.09	0.09			0.09
0.015	0	0.015	0.015	0.015		0.015
0	0	0		0.015	0.015	0.015
0	0	0		0.015	0.015	0.015
0	0	0		0.362		
0	0	0		0.362		
0	0	0		0.015	0.015	0.015
0	0	0		0.09	0.09	
0	0	0		0.09		0.09
0	0	0		0.09		0.09
0	0	0		0.09		0.09
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0			0.362	
0	0	0				0.362
0	0	0				0.362
0	0	0				0.362
0	0	0				0.362
0.015	0	0.105	0.105	1.144	3.755	1.868
			0.324	1.069	1.937	1.366

Dari tabel di atas, dapat dihitung nilai kemiripan dengan menggunakan Cosinus Similarity.
Dapat dilihat pada tabel 6, bahwa:

- Sudut yang dibentuk antara vektor dokumen D0 dan vektor dokumen D1 adalah 69,757 derajat (karena Cos 69,757 adalah 0,346). Maka dapat disimpulkan bahwa artikel D1 memiliki kesamaan isi sebesar 34,6% dengan artikel D0.

Tabel 6. Langkah Perhitungan 6

Cos(D0,D1)=	0.015/(0.324*1.069)=	0.346
Cos(D0,D2)=	0/(0.324*1.937)=	0
Cos(D0,D3)=	0.105/(0.324*1.366)=	0.442

- Sudut yang dibentuk antara vektor dokumen D0 dan vektor dokumen D2 adalah 90 derajat (karena Cos 90 adalah 0). Maka dapat disimpulkan bahwa artikel D2 tidak memiliki kesamaan dengan artikel D0.
- Sudut yang dibentuk antara vektor dokumen D0 dan vektor dokumen D3 adalah 63,768 derajat (karena Cos 63,768 adalah 0,442). Maka dapat disimpulkan bahwa artikel D3 memiliki kesamaan isi sebesar 44,2% dengan artikel D0.
- Dua vektor yang paling identik adalah D0 dan D3.

Jadi hasil dari percobaan diatas adalah artikel *Brazil yang menjadi tuan rumah piala dunia 2014, telah menyediakan logo untuk turnamen resmi berjudul "Inspiration",* memiliki tingkat kesamaan terbesar dengan dokumen : *Logo Turnamen* dibandingkan dengan dua artikel lainnya yaitu sebesar 0.442.

Urutan hasil kemiripan :

Tabel 7. Hasil Percobaan

1	2	3
D3	D1	D2

SIMPULAN DAN SARAN

Konsep *Semantic Similarity* dan metode *Cosinus Similarity* dapat diimplementasikan dalam pendeteksian kemiripan dua buah artikel web yang berbeda. Metode tersebut diimplementasikan dengan melakukan perhitungan tertentu untuk mendapatkan nilai kesamaan antara dua buah artikel tersebut. Dalam penelitian ini nilai cosinus tertinggi adalah 0.442 sehingga dapat disimpulkan artikel dengan nilai cosinus tersebut memiliki tingkat kesamaan terbesar dengan artikel penguji.Dengan aplikasi ini diharapkan dapat mengurangi tingkat plagiarisme yang dilakukan oleh para penulis akhir-akhir ini.

DAFTAR PUSTAKA

Adhit Herwansyah.2009. Aplikasi Pengkategorian Dokumen danPengukuran Tingkat Similaritas Dokumen Menggunakan Kata Kunci pada Dokumen Penulisan Ilmiah. Penulisan Akhir Jurusan Sistem Informasi Fakultas Ilmu Komputer & Teknologi Informasi Universitas Gunadarma.

Isa, Taufiq M, Taufik Fuadi Abidin. 2013. Mengukur Tingkat Kesamaan Paragraf Menggunakan Vector Space Model untuk Mendeteksi Plagiarisme.

Khairunnisa , Nova , Dadang Syarif and Ardianto Wibowo. “Aplikasi Pendeteksi Plagiat dengan Menggunakan Metode Latent Semantic Analysis”.

Santosa, Nugraha Imam, “Implementasi TF-IDF Untuk Pencarian Dokumen”.

Pradnyana, Gede Aditra, Arif Djunaidy. 2013. Metode Weighted Maximum Capturing untuk Klasterisasi Dokumen Berbasis Frequent Itemsets. Jurnal Ilmu Komputer Vol. 6 No.2