Jurnal Rekayasa Industri (JRI), Vol. 5 No. 2 Oktober 2023
p-ISSN: 2714-8882 / e-ISSN: 2714-8874

Penerapan Kemiripan Dokumen pada Mesin Pencar Menggunakan Metode Hellinger

Fatkhul Amin[1*], Eko Nur Wahyudi[2], Budi Hartono[3]
[1,2,3)] Fakultas Teknologi Informasi dan Industri, Program Studi Teknik Informatika, Universitas Stikubank Semarang
Jl. Tri Lomba Juang, Mugassari, Kec. Semarang Sel., Kota Semarang, Jawa Tengah 50241
Email: [1] fatkhulamin@edu.unisbank.ac.id, [2] eko@edu.unisbank.ac.id, [3] budihartono@edu.unisbank.ac.id

** Corresponding Author*

ABSTRAK

Penerapan alat pendeteksi kemiripan dokumen teks bahasa Indonesia dibuat untuk bisa menunjukkan seberapa tingkat pebedaan antar dokumen teks yang digunakan untuk mengurangi plagiarism. Model pendeteksi kemiripan dokumen menggunakan metode algoritma Hellinger yang ditujukan untuk menghasilkan tingkat akurasi yang tinggi. Proses Pra Perhitungan hellinger dilakukan tahap-tahap tokenisasi, penyaringan dan pembuatan akar kata. Penyaringan menggunakan *stopword* tala dan proses pembuatan akar kata menggunakan metode *rule base stemmer* bahasa Indonesia. Proses uji alat pendeteksi kemiripan didahului dengan uji persepsi menggunakan kamus besar bahasa indonesia untuk menetapkan kata yang dicari relevan atau tidak relevan. Hasil akhir pengujian alat pendeteksi kemiripan dokumen menggunakan metode Hellinger didapatkan tingkat akurasi rata-rata 0,71 dan tingkat rata-rata dokumen terambil 0,31.

Kata kunci: ***Hellinger,*** **Bahasa Indonesia, kemiripan dokumen**

ABSTRACT

The application of a similarity detection tool for Indonesian language text documents is made to be able to show the degree of difference between text documents that can be used to reduce plagiarism. The document similarity detection model uses the Hellinger algorithm method which is intended to produce a high level of accuracy. The Hellinger Pre-Calculation process is carried out in the stages of tokenization, filtering and making root words. Filtering using sopword tuning and the process of making word roots using the Indonesian language rule base stemmer method. The process of testing the similarity detection tool is preceded by a perception test using the Big Indonesian Dictionary to determine whether the word being searched is relevant or irrelevant. The final results of testing the document similarity detection tool using the Hellinger method obtained an average accuracy rate of 0.71 and an average retrieved document rate of 0.31.

Keywords: *Hellinger*, Bahasa Indonesia, kemiripan dokumen

I. PENDAHULUAN

Dalam sebuah dokumen terdapat banyak kata-kata yang telah tersusun dalam sebuah kalimat memiliki tingkat kemiripan antar dokumen, atau diantara banyak dokumen yang ada, tiap-tiap dokumen memiliki tingkat kemiripan. Kemiripan dokumen bisa dilihat dari hasil pencarian yang telah terlihat hasilnya. Mesin pencari bekerja sesuai perannya yaitu memberikan informasi yang relevan bagi penggunanya. Informasi yang dihasilkan oleh mesin pencari memiliki hasil yang berbeda dalam keluarannya karena algoritma yang berbeda pula.

Dokumentasi berupa teks bisa sangat bermanfaat jika diatur atau dikelola dalam sebuah *database* yang terintegrasi sehingga mudah menggunakannya. Penerapan dokumen yang terintegrasi belum banyak dilakukan oleh pelaku atau yang berkepentingan dengan data. Bahasa Indonesia dengan keunikannya tersendiri yaitu memiliki awalan, sisipan dan akhiran adalah obyek dokumen teks yang bisa digunakan untuk banyak hal seperti pengolahan data dan lain sebagainya. Jumlah data yang terus bertambah dari waktu ke waktu jika tidak dikelola dengan baik akan tidak bisa digunakan dengan maksimal penggunaannya. Data dengan isi atau konten yang sama akan terus berulang dan bertumpuk-tumpuk. Oleh karenanya diperlukan suatu pola penyimpanan data yang dibuat untuk mudah melakukan penyimpanan data dan mudah menemukan data. Model mesin pencari bisa digunakan untuk mendukung penyimpanan dan pengambilan kembali informasi yang efektif .

Mesin pencari dokumen perlu dibuat untuk mendukung penggunaan data dan sebagai alat pendeteksi kemiripan anat dokumen pada suatu sistem informasi atau pada suatu kumpulan data. Mesin pencari dibuat

dengan mempertimbangkan banyak hal seperti kategori dan isi yang bisa bermanfaat dalam kehidupan sehari hari, informasi dalam dokumen teks yang bisa diakses dengan cepat dan akurat. Penelitian terdahuluu tentang plagiarisme bisa dilihat pada tabel 1. Pada penelitian ini fokus kepada pembuatan mesin pencari untuk memeriksa plagiarisme pada dokumen teks menggunakan metode Hellinger.

Tabel 1. Penelitian Terdahulu

Austin J. Brockmeier, dkk (2017)	Arun S. Maiya (2014)	Shanmugasundaram Hariharan (2012)	Fatkhul Amin
Penelitian ini fokus pada pengukuran heterogenitas korelasi dan didefinisikan sebagai jarak antara matriks korelasi dan matriks korelasi dengan nilai konstan 0.	Penelitian fokus pada pendekatan yang efisien dan efektif untuk membangun dan melabeli jaringan tersebut. Model Visualisasi topik berdasarkan jaringan ini terbukti menjadi sarana yang kuat untuk mengeksplorasi, mencirikan, dan meringkas koleksi besar dokumen teks yang tidak terstruktur	Penelitian ini memfokuskan perhatiannya pada mengidentifikasi beberapa parameter kunci yang akan membantu mengidentifikasi plagiarisme dengan cara yang lebih baik	Penelitian fokus pada rancangan sebuah mesin pencari dokumen teks yang digunakan untuk mengukur atau melihat kemiripan antar dokumen yang ada pada sebuah database atau korpus.

Adapun solusi untuk mengatasi masalah ini adalah dengan mesin pencari menggunakan Metode *Hellinger* agar hasil pencarian informasi memiliki tingkat akurasi yang tinggi dan proses pencarian yang cepat.

II. METODE PENELITIAN

Pada Riset ini dirancang sebuah mesin pencari yang digunakan untuk mengukur atau melihat kemiripan antar dokumen yang ada pada sebuah *database*. Data yang digunakan didapatkan dari portal berita detik dot net atau informasi berita tentang dunia IT yang berfokus pada berita IT terbaru. Detiknet dipilih karena portal berita ini konsisten fokus di topik tentang IT dan berkesinambungan membagikan berita terkini tentang IT. Pada Riset ini akan dirancang sebuah mesin pencari yang digunakan untuk mengukur atau melihat kemiripan antar dokumen yang ada pada sebuah database. Berita di detiknet selanjutnya diambil dan disimpan dalam sebuah database yang digunakan untuk analisa kemiripan.

2.1 Alat Penyimpanan Dokumen Teks

Proses simpan database dilakukan dengan mengidentifikasikan terlebih dahulu dokumen yang ada atau membuat klaster dokumen yang ada untuk diberikan kode dokumen sehingga akan mudah dilakukan pengecekan lokasi dokumen ketika pengguna ingin mengetahuinya.

a. *Database* Korpus

Pada prinsipnya, setiap koleksi lebih dari satu teks dapat disebut dengan *corpus* (McEnery dan Wilson, 2001). istilah corpus dalam bahasa latin berarti body, maka corpus dapat didefinisikan sebagai isi setiap teks. Tapi istilah corpus ketika digunakan dalam konteks linguistic modern memiliki konotasi yang lebih spesifik. korpus merupakan referensi standar untuk berbagai bahasa yang diwakilinya. Hal ini mengandaikan ketersediaan yang luas kepada peneliti lain, Keuntungan dari korpus yang tersedia luas adalah akan memberikan tolak ukur yang dapat digunakan sebagai pembanding dalam studi.

2.2 Alat Pendeteksi Kemiripan Dokumen

Software ini dirancang dan dibuat untuk membantu menentukan kemiripan tiap dokumen dibandingkan dengan dokumen lain yang ada di korpus dengan model *queri*. Pendeteksi dibuat dengan model tampilan mesin pencari dan hasilnya akan dittampilkan berupa hasil pencarian yang tersusun berdasarkan pemeringkatan dengan metode Hellinger.

2.3 Metode *Hellinger*

Hellinger merupakan metode yang digunakan untuk menghitug tingkat kesamaan (*similarity*) antar dua buah objek. Untuk notasi himpunan dapat digunakan rumus (1):

$$= \left(2 \sqrt{1 - \sum_{i=1}^{d} \sqrt{P_i Q}\, i} \right) \quad (1)$$

dimana p dan q adalah dokumen yang berbeda. p_i adalah term i yang ada di dokurnen p q_i adalah term i yang ada di dokumen q.
Untuk menghitung kemiripan antara dua dokumen menggunakan metode Hellinger, langkah-langkah berikut dapat diikuti:

a. Pra-pemrosesan Dokumen: Lakukan pra-pemrosesan pada teks dokumen untuk membersihkan dan mempersiapkannya sebelum perhitungan. Langkah ini meliputi langkah-langkah seperti menghapus tanda baca, mengubah semua teks menjadi huruf kecil, menghilangkan kata-kata penghubung, dan melakukan *stemming* atau lemmatisasi jika diperlukan.
b. Representasi Vektor: Ubah teks dokumen menjadi representasi vektor. Salah satu metode yang umum digunakan adalah model ruang vektor atau bag-of-words (BoW). Dalam model ini, setiap dokumen direpresentasikan sebagai vektor di ruang berdimensi N, di mana N adalah jumlah kata unik dalam korpus. Setiap elemen vektor mewakili frekuensi kemunculan kata dalam dokumen.
c. Menghitung Histogram: Untuk masing-masing dokumen, hitung histogram dari representasi vektor. Histogram menggambarkan distribusi frekuensi kata-kata dalam dokumen.
d. Normalisasi Histogram: Normalisasi histogram dengan membagi setiap elemen dengan jumlah total elemen dalam histogram. Ini memastikan bahwa histogram mewakili distribusi probabilitas yang valid.
e. Menghitung Jarak Hellinger: Hitung jarak Hellinger antara dua histogram menggunakan rumus berikut:Hellinger_Distance = sqrt(0.5 * sum((sqrt(hist1[i]) - sqrt(hist2[i]))^2)). Di mana hist1 dan hist2 adalah histogram dua dokumen yang akan dibandingkan.Perhitungan ini melibatkan menghitung perbedaan antara akar kuadrat dari elemen-elemen histogram dan menghitung jarak akar kuadrat dari perbedaan tersebut.
f. Interpretasi Hasil: Semakin kecil nilai jarak Hellinger, semakin mirip dokumen-dokumen tersebu

2.4 Uji Metode *Hellinger*

Pendeteksi kemiripan dokumen menggunakan metode hellinger dipilih karena memiliki tingkat akurasi dokumen yang tinggi. Alat pendeteksi kemiripan dokumen selanjutnya dilakukan uji recall dan precission untuk mengetahui tingkat keakuran alat yang dibuat. Uji *recall* akan didapatkan informasi tentang banyaknya dokumen yang terambil ketika dilakukan pencarian dokumen. Uji *Precission* akan menghasilkan tingkat akurasi alat dari hasil pencarian. Proses pengujian ini menggunakan alat atau *tool* metode analisa persepsi dengan berdasarkan Kamus Besar Bahasa Indonesia.

III. HASIL DAN PEMBAHASAN

3.1 Implementasi Pendeteksi dokumen

a. Aplikasi Tampilan pendeteksi dokumen

Interface ini akan ditampilkan kolom *query* yang bisa digunakan untuk memasukkan *query* oleh pengguna. Model pendeteksi kemiripan dokumen didesain dengan tampilan mesin pencari. Tampilan yang sederhana dan mudah digunakan akan membuat pengguna mudah menggunakannya. Implementasi alat pendeteksi ini dilakukan dengan cara, pengguna menuliskan dokumen atau *copy* dokumen kedalam kotak pencarian. Selanjutnya setelah dokumen yang akan diperiksa kemiripannya sudah dimasukan, kemudian pengguna bisa klik tombol cari (Gambar 1).

.

Gambar 1. Tampilan *Home* Pendeteksi Kemiripan Dokumen

3.2 Implementasi Pendeteksi dokumen

a. Dokumentasi Dokumen kedalam korpus

Dokumen yang diambil dari portal berita detik dot net ditempatkan dalam database tersendiri dengan sebelumnya diberikan kode penomoran tiap dokumen. Input korpus dilakukan dengan cara memasukkan berita-berita di detik dot net Pengisian dokumen dikorpus dilakukan dengan sebelumnya dibuat kategorisasi berita dan diberikan penomoran. Pengelompokan topik diperlukan untuk membantu analisa, dan penomoran dokumen juga harus dilakukan untuk mengetahui posisi lokasi dokumen di korpus. Tabel 2 menunjukkan tampilan tabel korpus.

Tabel 2. Tabel Korpus

ID	Judul	Isi	Dokumen
1	PUBG jadi barang terlarang buat anak SD sebuah wilayah india	New delhi – main player uknowns battegrounds (PUBG) resmi dilarang di sekolah dasar daerah tertendtu di india. Wilayah gujarat, bagian barat india, yang melakukannya. Pekan ini departemen yang mengurusi pendidikan setara SD gujarat telah merilis sebuah surat edaran ke sekolah-sekolah diwilayah itu untu memastikan berjalannya larangan PUBG *MOBILE*.	GAM1
2	PUBG diibaratkan narkoba gara-gara nilai jeblok	Jakarta – popularitas player uknowns battegrounds (PUBG), secara khusus PUBG *mobile*, membuatnya dituding jadi biang keladi jebloknya nilai siswa disebuah wilayah di india. PUBG bahkan diibaratkan narkoba. Tudingan itu dilontarkan oleh asosiasi wali dan orang tua murid di jammu dan kasmir wilayah utara india.	GAM2

b. Tokenisasi

Tokenisasi adalah proses pemisahan teks atau kalimat menjadi unit yang lebih kecil, yang disebut "token." Token adalah unit dasar dalam pemrosesan bahasa alami (NLP), dan biasanya merupakan kata, frasa, atau simbol. Tokenisasi dilakukan pada tahap pertama proses pembuatan alat pendeteksi kemiripan dokumen. Tokenisasi dilakukan dengan pembuatan kode pemrograman yang akan menghasilkan pemisahan kalimat-kalimat menjadi kata yang selanjutnya kata tersebut dimasukkan ke dalam tabel-tabel. Tokenisasi menghasilkan kata dengan tampilan apa adanya, dalam bentuk huruf kecil semua. Semua kata yang telah dipisah dari kalimatnya selanjutnya disimpan dilengkapi dengan kode dokumen atau nomer dokumen. Tabel 3 menunjukkan hasil tokenisasi.

Tabel 3. Implementasi Proses Tokenisasi pada Tabel Hasil token

Judul	Term	Dokumen
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*New*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Delhi*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	-	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Main*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Battlegrounds*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*(Pubg)*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Mobile*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Resmi	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Dilarang	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Di	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Sekolah	GAM1

c. Penyaringan

Proses penyaringan merupakan proses penghapusan atau penyaringan elemen-elemen tertentu dari satu set data untuk mendapatkan subset yang memenuhi kriteria tertentu. Ini merupakan teknik yang umum digunakan dalam berbagai bidang, termasuk pengolahan bahasa alami, analisis data, pengolahan citra, dan banyak aplikasi lainnya. *Stopword* adalah kata atau kata-kata yang seringkali dianggap tidak memiliki makna penting dalam pemrosesan bahasa alami dan sering diabaikan atau dihilangkan dalam analisis teks. *Stopword* adalah kata-kata umum, seperti kata sambung, kata ganti, dan kata-kata umum lainnya yang tidak memberikan informasi yang signifikan tentang isi teks ketika dianalisis. Penghapusan *stopword* dari teks bertujuan untuk memfokuskan perhatian pada kata-kata kunci atau kata-kata yang lebih informatif. *Rule-based stemmer* adalah jenis algoritma dalam pemrosesan bahasa alami yang digunakan untuk mengubah

kata-kata dalam teks menjadi bentuk dasar atau "stem" mereka dengan mengikuti serangkaian aturan atau aturan gramatikal

Penyaringan dilakukan pada kata-kata yang sebelumnyanya telah dipisah-pisahkan pada proses token. Kata-kata selanjutnya dilakukan pencocokan atau diperiksa apakah ada kata yang termasuk dalam daftar kata *stopword*. Jika ada kata yang sama atau ada di *stopword*, maka sistem akan menghilangkan kata tersebut dan membuangnya dari data atau tabel data. Penyaringan menggunakan data *stopword* tala. Tabel 4 menunjukkan hasil proses penyaringan.

Tabel 4. Hasil Proses penyaringan

Judul	*Term*	Dokumen
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*New*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Delhi*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Main*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Main*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Battlegrounds*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	*Mobile*	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Resmi	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Dilarang	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Sekolah	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Dasar	GAM1
PUBG Jadi Barang Terlarang buat Anak SD Sebuah Wilayah India	Daerah	GAM1

d. *Stemming*

Pembuatan akar kata atau *stemming* dilakukan dengan pembuat kode pemrograman *stemming*. Pembuatan akar kata dilakukan dengan metode *rule base stemmer*. Sistem akan bekerja dengan cara menghilangkan awalan, sisipan dan akhiran. Sebuah kata jadian akan diproses menjadi kata dasar pada proses *stemming* ini. Tabel 5 menunjukkan hasil *stemmer*.

Tabel 5. Hasil *Stemmer*

Judul	*Term*	*Freq*	Freqpangkat
Dua kali obral iphone se, apa tujuan apple?	Jakarta	14	196
Dua kali obral iphone se, apa tujuan apple?	Apple	33	1089
Dua kali obral iphone se, apa tujuan apple?	Ketahuan	2	4
Dua kali obral iphone se, apa tujuan apple?	Masukan	1	1
Dua kali obral iphone se, apa tujuan apple?	Iphone	27	729
Dua kali obral iphone se, apa tujuan apple?	Laman	2	4
Dua kali obral iphone se, apa tujuan apple?	Sale	1	1
Dua kali obral iphone se, apa tujuan apple?	Resminya	2	4
Dua kali obral iphone se, apa tujuan apple?	Dipertahankan	1	1
Dua kali obral iphone se, apa tujuan apple?	Vendor	1	1
Dua kali obral iphone se, apa tujuan apple?	Ini	1	1

e. Hitung Hellinger

Proses perhitungan metode hellinger dilakukan pada proses pencarian dokumen mirip. Proses tokenisasi, penyaringan dan *stemming* menyiapkan data berupa hasil akhir kumpulan kata dasar, selanjutnya akan dilakukan perhitungan-perhitungan dimulai dari *indexing*, hitung tfidf, hitung bobot dan hitung similaritas hellinger. Perhitungan dibuat menjadi kode pemrograman yang bisa dieksekusi menggunakan sarana laptop dan *smartphone*.

3.3 Studi Kasus *Keyword* Bahasa Indonesia

Studi kasus pada aplikasi mesin pencari ini menggunakan dokumen-dokumen teks berbahasa jawa krama. *Query* yang dimasukkan pada mesin pencari adalah *keyword* dengan 2 *term* yaitu "*game* apple", "*game* mark", "*game* facebook" , 3 *term* "rencana *game* instagram", "pemblokiran *game* di facebook", "Apple dan iphone", 4 *term* "*Hoaks game* di instagram", "*Qualcomm* dan CEO apple". 5 *term* "facebook instagram dan whatsup akuisisi" Gambar 2 menunjukkan contoh implementasi Pendeteksi kemiripan dokumen.

Gambar 2. Implementasi Pendeteksi dokumen

3.4 *Recall* dan *Precision*

Nilai *recall* dihitung menggunakan persamaan (2)

$$R = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection} \quad (2)$$

Dengan R adalah *recall,*maka nilai *R* didapatkan dengan membandingkan *Number of relevant items retrieved* dengan *Total number of relevant items in collection*. *Recall* adalah dokumen yang terpanggil dari Pendeteksi kemiripan dokumen sesuai dengan permintaan *user* yang mengikuti pola dari Pendeteksi kemiripan dokumen. Nilai *recall* makin besar belum cukup untuk menilai suatu Pendeteksi kemiripan dokumen baik atau tidak.

Nilai *precision* dihitung menggunakan persamaan (3)

$$P = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \quad (3)$$

Dengan *P* adalah *Precision.* maka nilai *P* didapatkan dengan membandingkan *Number of relevant items retrieved* dengan *Total number of items retrieved.* Precision adalah jumlah dokumen yang terpanggil dari *database* relevan setelah dinilai user dengan informasi yang dibutuhkan. Semakin besar nilai *precision* suatu Pendeteksi kemiripan dokumen, maka Pendeteksi kemiripan dokumen dapat dikatakan baik.

3.5 Uji *recall* dan *precision*

a. Deklarasi Pengujian keyword menggunakan Persepsi

Deklarasi uji penting dilakukan untuk memastikan apakah dokumen yang didapatkan oleh mesin pendeteksi bisa dikatakan relevan atau tidak relevan. Dokumen dikatakan relevan jika arti kata yang dicari sesuai dengan pendeklarasian diawal. Dokumen dikatakan tidak relevan jika tidak sesuai dengan kata yang telah dideklarasikan sebelumnya. Pendeklarasian ini penting dilakukan untuk bisa menghasilakn nilai uji recall dan precission.

3.5.1 Perhitungan Recall dan Precision

Hasil perhitungan Recall untuk keyword “rencana *game* instagram” adalah sebagai berikut;

$$R = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection}$$

Recall = 0.47

Hasil perhitungan Precision untuk keyword “wayang jawi” adalah sebagai berikut;

$$P = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved}$$

Precision = 0,89

Hasil lengkap bisa dilihat pada tabel 5. Hasil perhitungan rata-rata untuk Recall dan precision adalah sebagai berikut;

Rata-rata Recall = 0,31

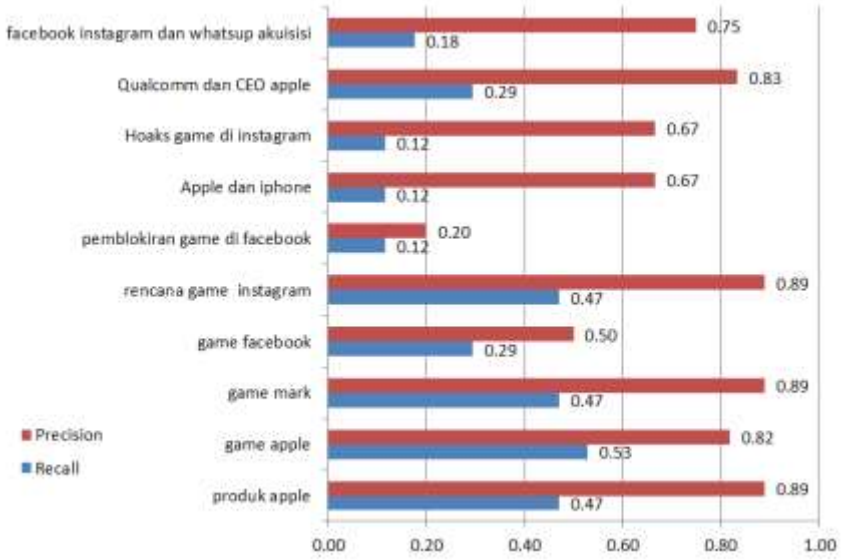
Rata-rata Precision = 0,71

Tabel 5. Hasil Pengujian *Recall* dan *Precision*

No	*Query*	*Recall*	*Precision*
1	Produk apple	0,47	0,89
2	*Game* apple	0,53	0,82
3	*Game* mark	0,47	0,89
4	*Game* facebook	0,29	0,50
5	Rencana *game* instagram	0,47	0,89
6	Pemblokiran *game* di facebook	0,12	0,20
7	Apple dan iphone	0,12	0,67
8	Hoaks *game* di instagram	0,12	0,67
9	Qualcomm dan CEO apple	0,29	0,83
10	Facebook instagram dan whatup akuisisi	0,18	0,72

3.6 Diagram Hasil Uji Recall dan Precision

Diagram hasil uji *recall* dan *precision* bisa dilihat pada gambar 3



Gambar 3. Diagram Hasil perhitungan *Recall* dan *Precission*

IV. SIMPULAN

Pendeteksi kemiripan dokumen Metode Hellinger dibangun memiliki keunggulan mampu melakukan pendeteksian dokumen dengan hasil kemiripan dokumen yang akurat (*precission* = 0,89). Hasil Uji *recall* dan *precision* pendeteksi kemiripan dokumen Metode Hellinger menunjukkan hasil pencarian dokumen teks memiliki rata-rata *recall* = 0,31 dan rata-rata *precision* = 0,71.

DAFTAR PUSTAKA

Arslan, A., & Velioglu, S. G. (2015). The Comparison of Hellinger and Kullback-Leibler Divergences on Fuzzy Information Measure. *Journal of Intelligent & Fuzzy Systems*, Vol. 29(6), 2421-2428.

Arman, M. (2020). Metode pertahanan web server terhadap distributed slow HTTP DoS attack. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, Vol. 7(1), 56-70

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, Vol. 24, 136-158.

Fiedor, P., & Kawałek, A. (2019). A new distance measure based on Hellinger distance and fuzzy numbers for outlier detection. *Central European Journal of Operations Research*, Vol. 27(3), 783-799.

Karlis, D., & Xekalaki, E. (1998). Minimum Hellinger distance estimation for Poisson mixtures. *Computational Statistics & Data Analysis*, Vol. 29(1), 81-103.

Kraljeta, V. (2012). Business Constellations-New Tool for Entrepreneurial Learning. *Učenje za poduzetništvo*, Vol. 2(2), 177-187.

Kumar, N., Kaur, A., & Arora, N. (2019). A comparative analysis of Jensen-Shannon and Hellinger distances for document clustering. *International Journal of Applied Engineering Research*, Vol. 14(12), 3218-3221.

Lee, C. H. (2007). A Hellinger-based discretization method for numeric attributes in classification learning. *Knowledge-Based Systems*, Vol. 20(4), 419-425.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The annals of statistics*, Vol. 22(2), 1081-1114.

Li, H., Liu, Y., & Liu, Y. (2018). Application of Hellinger Distance in Image Retrieval. Journal of Physics: Conference Series, 1049(1), 012046.

Lourenzutti, R., & Krohling, R. A. (2014). The Hellinger distance in Multicriteria Decision Making: An illustration to the TOPSIS and TODIM methods. *Expert Systems with Applications*, Vol. 41(9), 4414-4421.

Madani, A., & Ahmadi, A. (2021). Performance Comparison of Hellinger Distance and Entropy on Particle Swarm Optimization for Data Clustering. *Journal of Computational and Theoretical Nanoscience*, Vol. 18(9), 4999-5004

Ramona, S. A. M., Pompiliu, C. M., & Constantin, S. L. (2017). Attainment of K-means algorithm using hellinger distance. *Economic Sciences Series*, Vol. 17(2), 324-329.

Salvi, A., & Sathya, G. (2020). Hellinger Distance Based Fuzzy Clustering for Medical Image Segmentation. *Journal of King Saud University-Computer and Information Sciences*, Vol. 32(1), 31-39.

Wahyudin, W. (2020). Aplikasi Topic Modeling Pada Pemberitaan Portal Berita Online Selama Masa Psbb Pertama. In *Seminar Nasional Official Statistics* Vol. 2020(1), 309-318.