

Implementasi Metode Cosine Similarity Untuk Deteksi Kemiripan Data Pada Dokumen Word Dan Perbedaan Isi Dokumen Word

Implementation Of The Cosine Similarity Method For Detecting Data Similarity In Word Documents And Differences In The Contents Of Word Documents

Safriandi Hatoguan Sihombing^{a,1}, Mas Ayoe Elhias Nasution^{b,2}

^aMahasiswa Universitas Potensi Utama, JL. K. L. Yos Sudarso KM 6,5 No. 3-A, Medan, 20241, Indonesia

^bDosen Universitas Potensi Utama, JL. K. L. Yos Sudarso KM 6,5 No. 3-A, Medan, 20241, Indonesia

Email: safriandi.sssw@gmail.com¹, khairan.adwa@gmail.com²

ABSTRAK

Sebuah dokumen juga berisi tentang teks yang mudah dipahami dan disalin oleh orang-orang yang mendapatkan sebuah dokumen diberbagai tempat misalnya internet. Dokumen digunakan untuk membentuk sebuah laporan yang dapat digunakan untuk keperluan penelitian. Masalah yang terjadi adalah dokumen banyak digunakan dengan isi yang sama, sehingga hal ini membuktikan bahwa dokumen yang digunakan bukan hasil pemikiran sendiri akan tetapi menggunakan dokumen orang lain untuk kepentingannya sendiri. Disisi lain dokumen word mengandung kata kata yang sama sehingga sulit mendeteksinya hanya dengan melihat. Oleh karena itu dibutuhkan sebuah cara yang dapat mendeteksi kemiripan dan perbedaan isi dokumen word sehingga seseorang dapat mengenali keaslian sebuah dokumen yang ditulis berdasarkan pemikiran sendiri dan dibutuhkan cara yang dapat mendeteksi isi dokumen word sehingga dapat mengenali kesamaan dan perbedaan isinya. Penelitian ini membuat aplikasi yang dapat mendeteksi isi dokumen word. Dengan adanya aplikasi implementasi metode cosine similarity untuk deteksi kemiripan dan perbedaan isi dokumen word maka dapat dengan mudah mendeteksi kemiripan dan perbedaan sebuah dokumen word dan mendapat kemudahan dalam pendeteksian dokumen word.

Kata Kunci: Implementasi, Metode Cosine Similarity, Deteksi, Kemiripan, Perbedaan, Dokumen Word.

ABSTRACT

A document also contains text that is easy to understand and copy by people who get a document in various places, for example the internet. Documents are used to form a report that can be used for research purposes. The problem that occurs is that many documents are used with the same content, so this proves that the document used is not the result of one's own thoughts but uses other people's documents for one's own interests. On the other hand, Word documents contain the same words so it is difficult to detect them just by looking. Therefore, we need a method that can detect similarities and differences in the contents of a word document so that someone can recognize the authenticity of a document written based on their own thoughts and we need a method that can detect the contents of a word document so that they can recognize the similarities and differences in its contents. This research creates an application that can detect the contents of a Word document. With the implementation of the cosine similarity method for detecting similarities and differences in the contents of word documents, it is possible to easily detect similarities and differences in a word document and make it easier to detect word documents.

Keywords : Implementation, Cosine Similarity Method, Detection, Similarity, Difference, Word Document.

Info Artikel :

Disubmit: 10 Februari 2024

Direview: 06 Maret 2024

Diterima : 04 April 2024

Copyright © 2024 - Journal UPU. All rights reserved..

1. PENDAHULUAN

Teks merupakan seperangkat tanda yang ditransmisikan dari seorang pengirim kepada seorang penerima melalui medium atau kode-kode tertentu. Teks hampir sama dengan wacana, hanya saja teks hanya dapat disampaikan dalam bentuk tulisan saja, sedangkan wacana dapat disampaikan dalam bentuk lisan maupun tertulis. [1][2]. Sebuah dokumen juga berisi tentang teks yang mudah dipahami dan disalin oleh orang-orang yang mendapatkan sebuah dokumen diberbagai tempat misalnya internet. Dokumen digunakan untuk membentuk sebuah laporan yang dapat digunakan untuk keperluan penelitian. Masalah yang terjadi adalah dokumen banyak digunakan dengan isi yang sama, sehingga hal ini membuktikan bahwa dokumen yang digunakan bukan hasil pemikiran sendiri akan tetapi menggunakan dokumen orang lain untuk kepentingannya sendiri. Disisi lain dokumen word mengandung kata kata yang sama sehingga sulit mendeteksinya hanya dengan melihat. Dokumen adalah kumpulan berkas berisi teks yang berupa hasil olahan. [3][4]. MS Word adalah aplikasi buatan microsoft yang biasa digunakan untuk mengelola dan membuat berkas. [5][6]. Oleh karena itu dibutuhkan sebuah cara yang dapat mendeteksi kemiripan dan perbedaan isi dokumen word sehingga seseorang dapat mengenali keaslian sebuah dokumen yang ditulis berdasarkan pemikiran sendiri dan dibutuhkan cara yang dapat mendeteksi isi dokumen word sehingga dapat mengenali kesamaan dan perbedaan isinya.

Penggunaan komputer sudah banyak membantu kinerja manusia dalam berbagai hal terutama pengelolaan data dan informasi. Oleh karena itu peneliti memanfaatkan teknologi komputer untuk membuat sebuah sistem yang dapat mendeteksi kemiripan gambar. Dalam menggunakan sistem deteksi dibutuhkan metode yang dapat memberikan hasil kemiripan yang baik dan tepat.

Berdasarkan penelitian yang dilakukan oleh Herlambang, dkk (2021) mengenai Analisa Dan Perancangan Sistem Pendeteksi Plagiarisme Skripsi Pada Stmik Insan Pembangunan Menggunakan Metode *Cosine Similarity*, Herlambang, dkk menggunakan metode *Cosine Similarity* untuk deteksi plagiarisme skripsi.

Berdasarkan penelitian yang dilakukan oleh Kambey, dkk (2020) mengenai Penerapan *Clustering* pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia, Kambey, dkk menggunakan metode *Cosine Similarity* untuk deteksi dokumen teks bahasa indonesia.

Berdasarkan penelitian yang dilakukan oleh Nugroho, dkk (2020) mengenai Penerapan Algoritma *Cosine Similarity* untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat, Nugroho, dkk menggunakan metode *Cosine Similarity* untuk deteksi kesamaan konten pada sistem informasi penelitian dan pengabdian kepada masyarakat.

Dari beberapa penelitian terdahulu yang menggunakan metode *Cosine Similarity* untuk mengatasi berbagai masalah kemiripan dan perbedaan dokumen sehingga memberikan hasil yang baik dan tepat, maka peneliti menggunakan *Cosine Similarity* untuk mendeteksi kemiripan dan perbedaan isi dokumen word. Metode *Cosine Similarity* adalah algoritma yang dapat digunakan untuk melakukan perhitungan kesamaan dari sebuah dokumen. [7][8]. Dengan adanya aplikasi implementasi metode *cosine similarity* untuk deteksi kemiripan dan perbedaan isi dokumen word maka dapat dengan mudah mendeteksi kemiripan dan perbedaan sebuah dokumen word dan mendapat kemudahan dalam pendeteksian dokumen word.

2. METODE

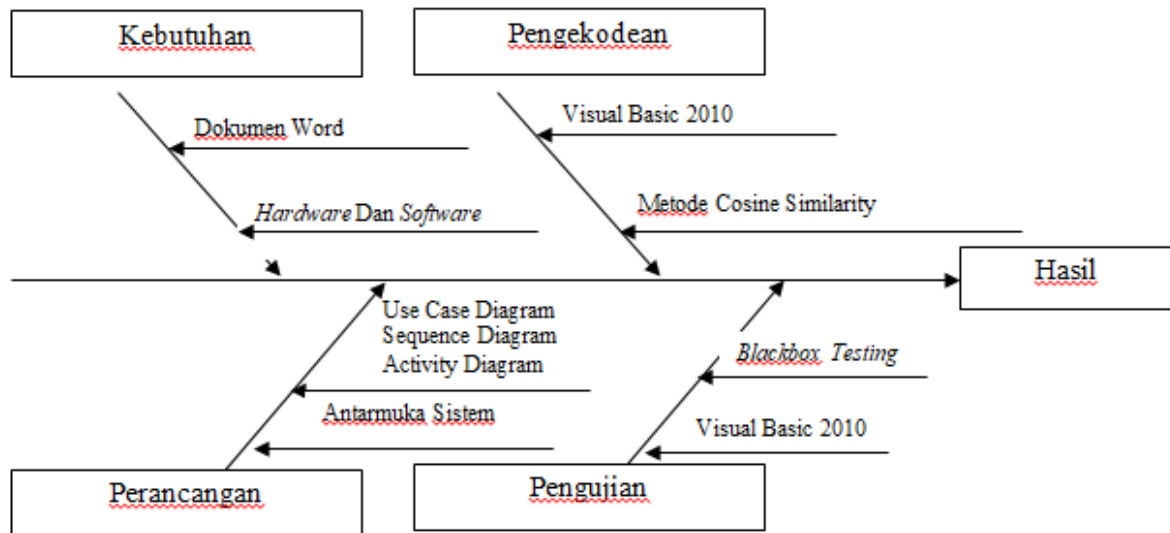
A. Metodologi Penelitian

Pengumpulan data yang peneliti lakukan menggunakan beberapa teknik sebagai berikut:

1. Studi Kelapangan
 - a. Tanya Jawab

- Peneliti melakukan tanya jawab mengenai penelitian kepada ahli untuk mendapatkan kejelasan mengenai deteksi kemiripan data dokumen word.
- b. Sampel Data
Peneliti mengambil beberapa contoh penelitian terdahulu dan data-data dari tempat riset untuk digunakan sebagai contoh untuk penelitian ini.
2. Studi Kepustakaan
Studi kepustakaan yang peneliti yaitu mengutip beberapa jurnal yang berkaitan dengan deteksi kemiripan data dokumen word.

Peneliti memodelkan metodologi penelitian menggunakan diagram *Fish Bone*. Beberapa tahapan yang digunakan dalam penelitian ini adalah sebagai berikut:



Gambar 1. Diagram *Fish Bone* Metodologi Penelitian

Keterangan:

1. Kebutuhan
Peneliti mengumpulkan data-data untuk kebutuhan penelitian yaitu dokumen word, hardware dan software yang digunakan untuk penelitian ini.
2. Perancangan
Peneliti menggunakan pemodelan UML yaitu *use case diagram*, *activity diagram* dan *sequence diagram* untuk perancangan sistem.
3. Pengkodean
Peneliti menggunakan pemrograman *Visual Basic 2010*.
4. Pengujian
Pengujian program menggunakan *Visual Basic 2010* dan pengujian teori menggunakan *blackbox testing*.
5. Hasil
Pada tahapan ini peneliti telah menyelesaikan seluruh penelitian baik teori maupun aplikasi yaitu Aplikasi Implementasi Metode *Cosine Similarity* Untuk Deteksi Kemiripan Data Pada Dokumen Word.

B. Metode Consine Similarity

Metode *Consine Similarity* adalah algoritma yang dapat digunakan untuk melakukan perhitungan kesamaan dari sebuah dokumen. Untuk notasi himpunan pada *consine similarity* digunakan rumus:

$$CosSim \propto \frac{A.B}{|A|.|B|} + \frac{\sum_{i=1}^n A_i X B_i}{\sqrt{\sum_{i=1}^n (B_i)^2} \times \sqrt{\sum_{i=1}^n (A_i)^2}} \dots\dots\dots(1)$$

Dimana:

A = Vektor A, yang akan dibandingkan kemiripannya

B = Vektor B, yang akan dibandingkan kemiripannya

$A.B$ = dot product antara vektor A dan vektor B

$|A|$ = panjang vektor A

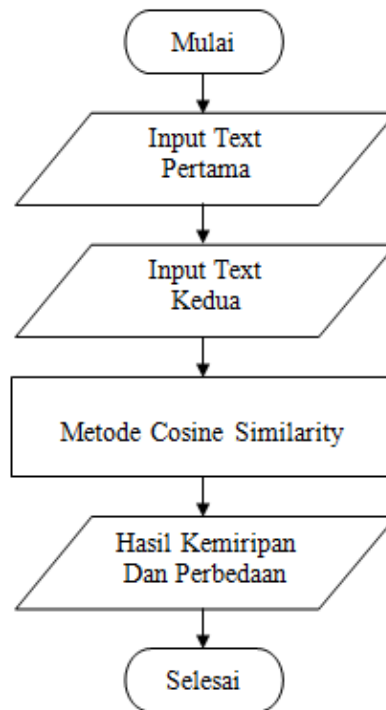
$|B|$ = panjang vektor B

$|A||B|$ = cross product antara $|A|$ dan $|B|$

Langkah-langkah perhitungan manual algoritma Cosine Similarity antara lain:

1. Langkah pertama terlebih dahulu kita tentukan *query* dari jawaban (A) dan *query* dari *key* jawaban (B) dan kemudian gabungan dari kedua *query* tersebut (*Queries*).
2. Dari ketiga *query* yang ada selanjutnya dilakukan penghapusan *stoplist* atau tanda baca yang tidak mempengaruhi penilaian, seperti tanda titik(.), tanda koma(,), tanda seru(!), dan tanda baca lainnya.
3. Menghilangkan *stopword* (kata-kata umum) yang lazim digunakan pada ketiga *query* tersebut kata-kata umum dalam suatu *query*, seperti "di", "dan", "tetapi", "jika", "namun", dan kata hubung lainnya.
4. Menghitung nilai *term frequency* dari *query* jawaban dan *query key* jawaban terhadap *queries*.
5. Hitung nilai banyaknya *file* (N) atau *document frequency* (n) yang memiliki suatu *term* untuk tiap *term* dalam *queries*. f)
6. Menghitung nilai *inverse document frequency* dengan rumus $\log (N/n) + 1$. [9][10].

Flowchart metode cosine similarity:



Gambar 2. Flowchart Metode Cosine Similarity

Contoh:

Jika terdapat kalimat berikut:

A: Mengajar di Universitas Potensi Utama

B: Belajar di Universitas Potensi Utama

Kemudian:

n_A = jumlah kemunculan kata indeks ke-n dari daftar kata pada kalimat A.

nB = jumlah kemunculan kata indeks ke-n dari daftar kata pada kalimat B.

Tabel 1. Uji Kemiripan Dan Perbedaan

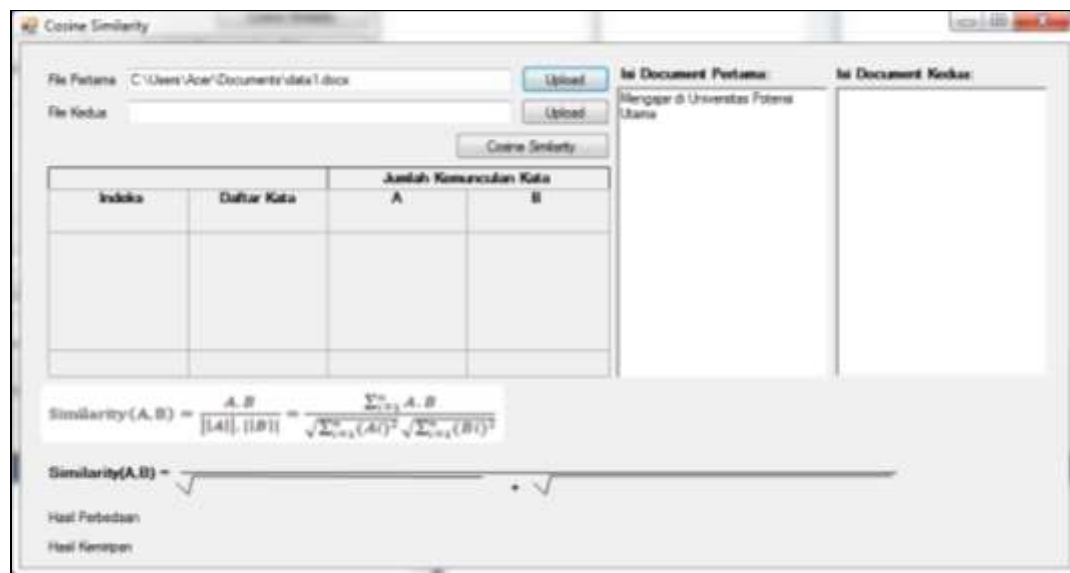
Indeks	Daftar Kata	Jumlah Kemunculan Kata	
		A	B
1	Mengajar	1	0
2	di	2	2
3	Universitas	2	2
4	Potensi	2	2
5	Utama	2	2

Kemudian, untuk mencari tingkat kemiripan menggunakan rumus cosine similarity:
Kemiripan text yang diperoleh yaitu sebesar 94%.
Maka perbedaan sebesar 6%.

3. HASIL DAN PEMBAHASAN

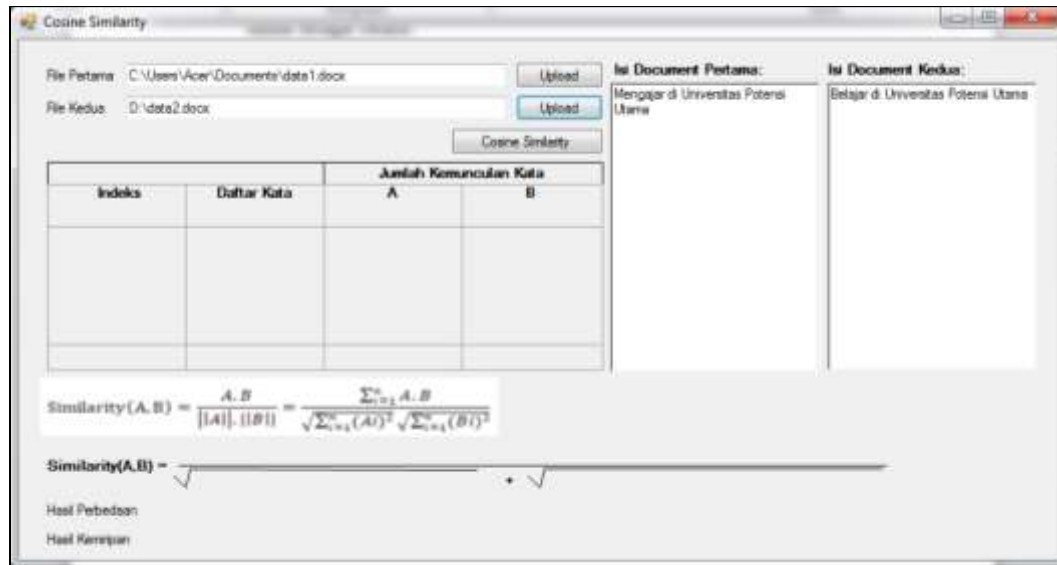
A. Hasil

Tampilan hasil dari aplikasi Implementasi Metode *Cosine Similarity* Untuk Deteksi Kemiripan Data Pada Dokumen *Word* dapat dilihat sebagai berikut:



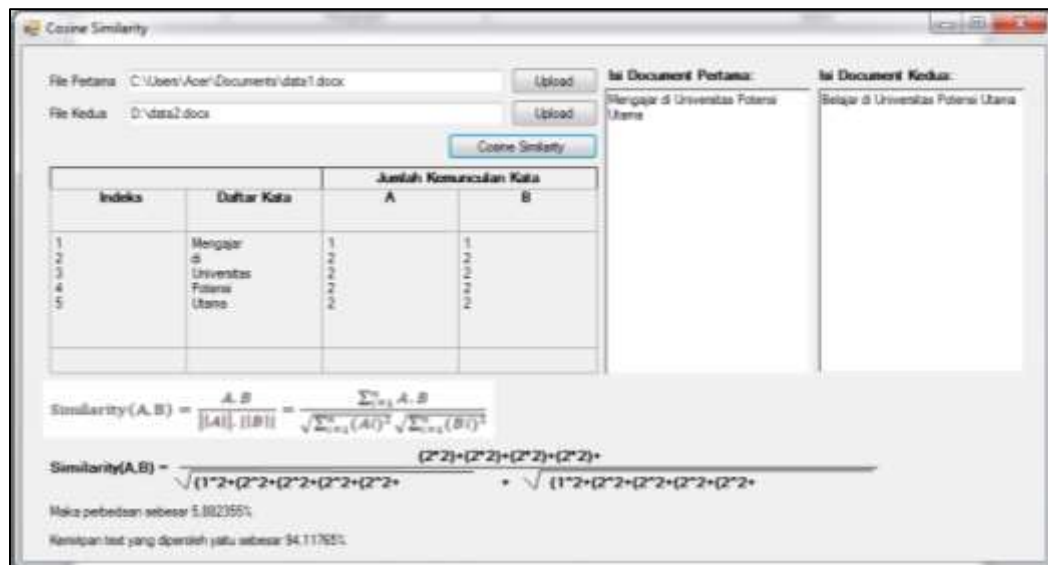
Gambar 3. Tampilan Upload File Pertama

Pada Gambar 3 merupakan tahapan awal upload file pertama dan isi file dapat dilihat pada *richtext* kanan pada aplikasi. Kemudian tahap selanjutnya adalah upload file kedua dan dapat dilihat pada Gambar 4.



Gambar 4. Tampilan Upload File Kedua

Pada Gambar 4 merupakan tahapan awal upload file kedua dan isi file dapat dilihat pada *richtextbox* kanan pada aplikasi. Kemudian tahap selanjutnya adalah menampilkan hasil kemudian mengklik tombol *Cosine Similarity* dan dapat dilihat pada Gambar 5.

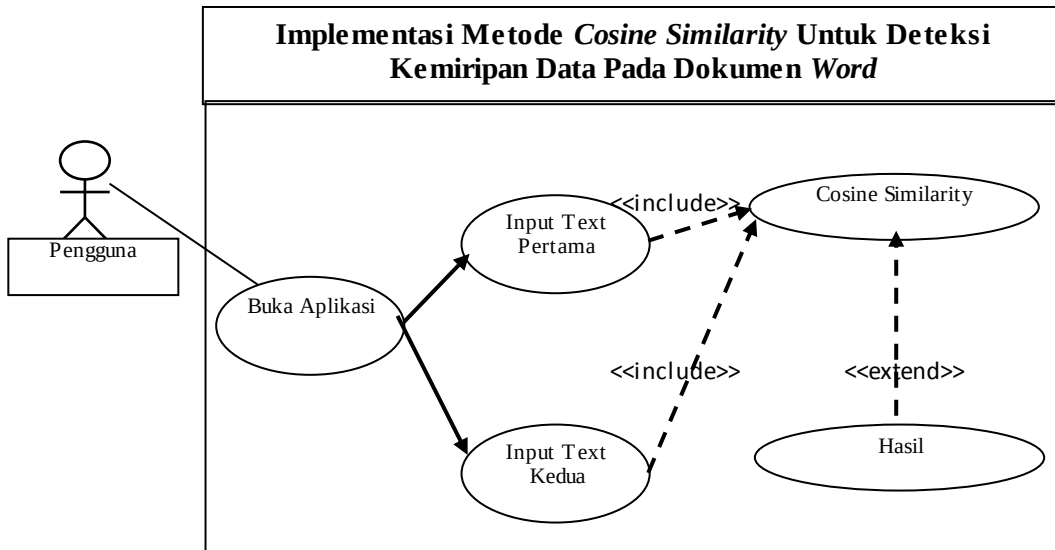


Gambar 5. Tampilan Hasil Metode Cosine Similarity

B. Pembahasan

1. Use Case Diagram

Perancangan dimulai dari identifikasi aktor dan bagaimana hubungan antara aktor dan use case didalam sistem. Perancangan *Use Case Diagram* dapat dilihat pada gambar 6.



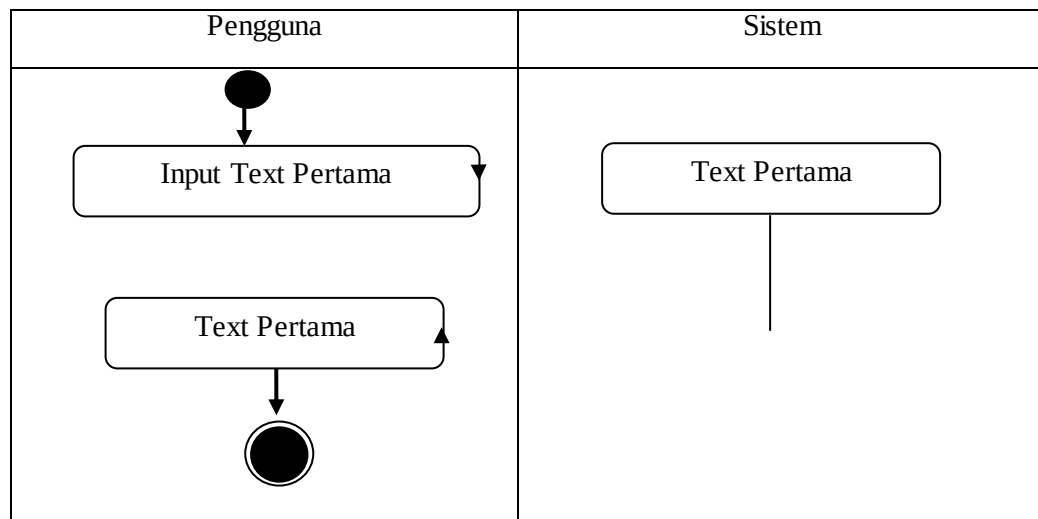
Gambar 6. Use Case Diagram Implementasi Metode Cosine Similarity Untuk Deteksi Kemiripan Data Pada Dokumen Word

2. Activity Diagram

Serangkaian kerja melakukan deteksi kemiripan dan perbedaan data dokumen word dapat terlihat sebagai berikut:

a. Activity Diagram Input Text Pertama

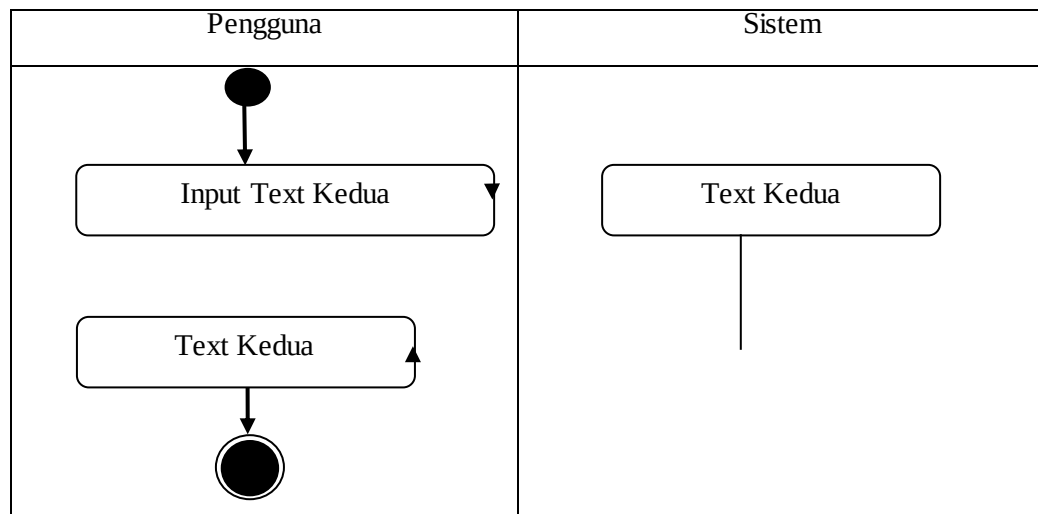
Gambar 7 menjelaskan Pengguna aplikasi yang melakukan *input text* pertama untuk deteksi kemiripan dan perbedaan data dokumen word:



Gambar 7. Activity Diagram Input Text Pertama

b. Activity Diagram Input Text Kedua

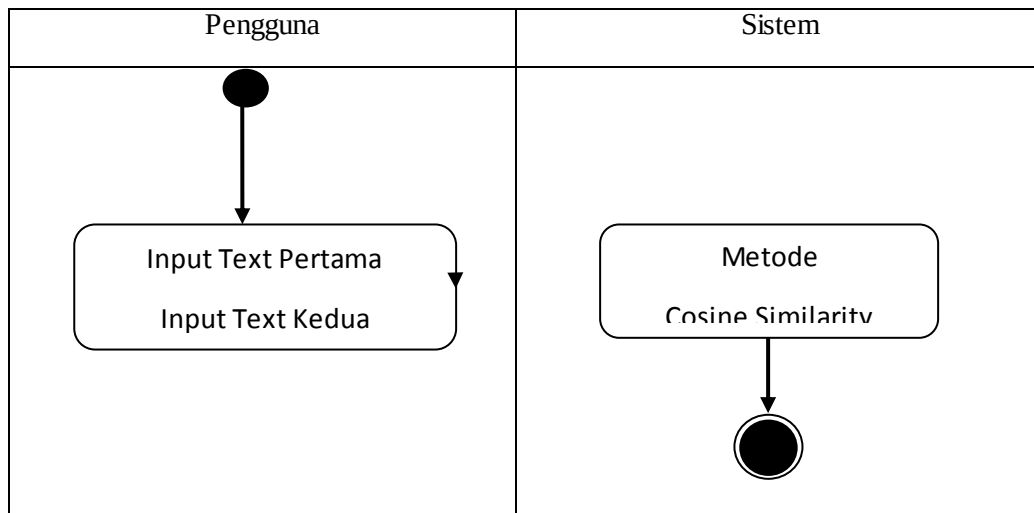
Gambar 8 menjelaskan Pengguna aplikasi yang melakukan *input text* kedua untuk deteksi kemiripan dan perbedaan data dokumen word:



Gambar 8. Activity Diagram Input Text Kedua

b. Activity Diagram Cosine Similarity

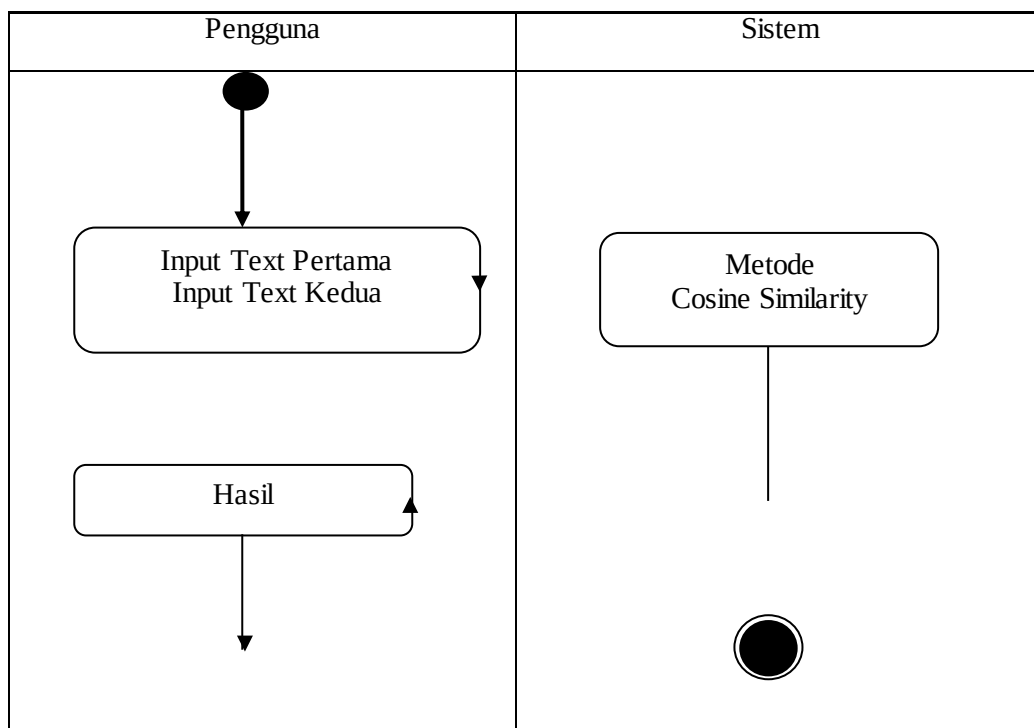
Gambar 9 menjelaskan Pengguna aplikasi yang menggunakan metode *Cosine Similarity* untuk deteksi kemiripan dan perbedaan data dokumen word:



Gambar 9. Activity Diagram Cosine Similarity

c. Activity Diagram Hasil

Gambar 10 menjelaskan Pengguna aplikasi yang mendapatkan hasil dari deteksi kemiripan dan perbedaan data dokumen word:



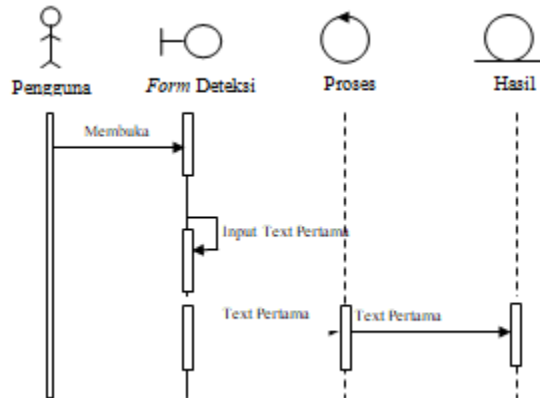
Gambar 10. Activity Diagram Hasil

3. Sequence Diagram

Aktivitas yang dilakukan untuk melakukan deteksi kemiripan dan perbedaan data dokumen word dapat dilihat sebagai berikut:

1. Sequence Diagram Input Text Pertama

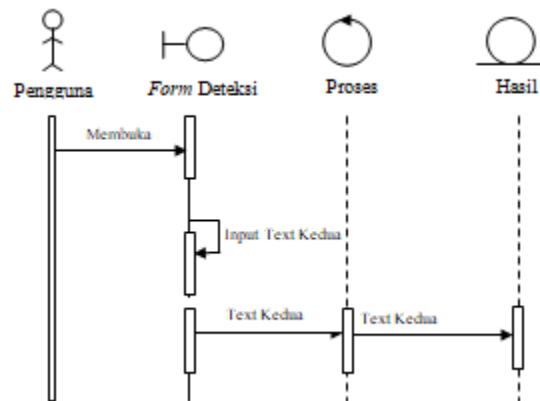
Gambar 11 menjelaskan Pengguna aplikasi yang melakukan *input text* pertama untuk deteksi kemiripan dan perbedaan data dokumen *word*:



Gambar 11. Sequence Diagram Input Text Pertama

2. Sequence Diagram Input Text Kedua

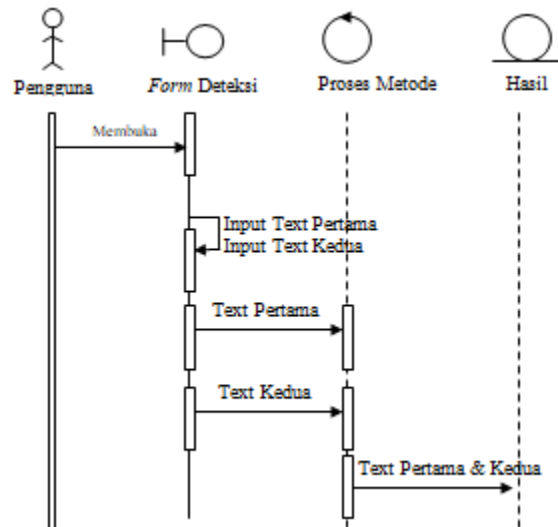
Gambar 12 menjelaskan Pengguna aplikasi yang melakukan *input text* kedua untuk deteksi kemiripan dan perbedaan data dokumen *word*:



Gambar 12. Sequence Diagram Input Text Kedua

3. Sequence Diagram Metode

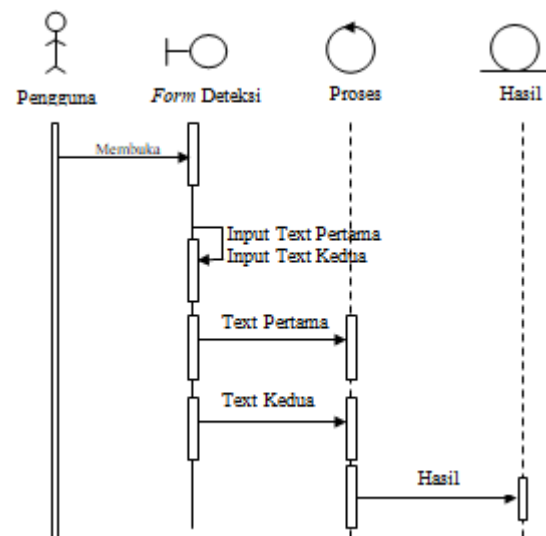
Gambar 13 menjelaskan Pengguna aplikasi yang melakukan *input text* pertama dan *text* kedua untuk deteksi kemiripan dan perbedaan data dokumen *word*:



Gambar 13. Sequence Diagram Metode

4. Sequence Diagram Hasil

Gambar 14 menjelaskan Pengguna aplikasi yang melakukan *input text* pertama dan kedua untuk deteksi kemiripan dan perbedaan data dokumen *word*:



Gambar 14. Sequence Diagram Hasil

Untuk membuat aplikasi Implementasi Metode *Cosine Similarity* Untuk Deteksi Kemiripan Data Pada Dokumen *Word* dapat dibutuhkan sebuah perangkat keras dan perangkat lunak agar pembuatan aplikasi dapat berjalan dengan baik. Kebutuhan perangkat keras dan perangkat lunak untuk membuat aplikasi adalah sebagai berikut:

1. Satu unit laptop dengan spesifikasi sebagai berikut:
 - a. *Processor* minimal *Core 2 Duo*
 - b. *RAM* minimal 1 Gb
 - c. *Hardisk* minimal 80 Gb
2. Perangkat Lunak dengan spesifikasi sebagai berikut:
 - a. *Sistem Operasi Windows*
 - b. *Visual Basic 2010*

3. Uji Coba Program

Uji coba terhadap aplikasi bertujuan untuk memastikan bahwa aplikasi sudah berada pada kondisi siap pakai. Instrumen yang digunakan untuk melakukan pengujian ini yaitu dengan menggunakan *Blackbox Testing*:

Tabel 2. Blackbox Testing Form Deteksi

No	Form Deteksi	Keterangan	Validitas
1.	Klik Tombol Upload Gambar Pertama	Aplikasi menampilkan pemilihan file	Valid
2.	Klik Tombol Upload Gambar Kedua	Aplikasi menampilkan pemilihan file	Valid
3.	Klik Tombol Metode <i>Cosine Similarity</i>	Aplikasi menampilkan hasil perhitungan metode <i>Cosine Similarity</i>	Valid

4. Hasil Uji Coba

Setelah melakukan uji coba terhadap sistem, maka dapat disimpulkan hasil yang didapatkan yaitu:

Tabel 3. Hasil Uji Di Teori Dan Aplikasi

Indeks	Daftar Kata	Jumlah Kemunculan Kata Secara Teori		Jumlah Kemunculan Kata Secara Aplikasi	
		A	B	A	B
1	Mengajar	1	0	1	0
2	di	2	2	2	2
3	Universitas	2	2	2	2
4	Potensi	2	2	2	2
5	Utama	2	2	2	2
6	Belajar	0	1	0	1
Persentase		94%		94,11765%	

- Interface* rancangan telah sesuai dengan *Interface* hasil.
- Metode *Cosine Similarity* telah diterapkan pada aplikasi yang dibuat.
- Interface* aplikasi bersifat *user friendly* sehingga pengguna dapat menggunakannya dengan mudah.
- Aplikasi yang telah dibuat berjalan dengan baik.
- Aplikasi yang telah dibuat tidak memiliki kesalahan logika.

4. KESIMPULAN

Berdasarkan pembahasan dari bab-bab sebelumnya yang telah dilakukan maka dapat diambil beberapa kesimpulan sebagai berikut:

- Dengan adanya aplikasi Implementasi Metode *Cosine Similarity* Untuk Deteksi Kemiripan Data Pada Dokumen *Word* maka dapat mendeteksi kemiripan dan perbedaan isi dokumen *word*.
- Dengan menggunakan file *word* dan memanfaatkan isi teks didalamnya maka dapat mengimplementasikan metode *cosine similarity* untuk deteksi kemiripan data dokumen *word*.
- Dengan menggunakan pemrograman *visual basic* 2010 maka dapat menghasilkan aplikasi Implementasi Metode *Cosine Similarity* Untuk Deteksi Kemiripan Data Pada Dokumen *Word*.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Universitas Potensi Utama yang telah memberikan kesempatan pada penulis agar menyelesaikan karya ilmiah ini. Penulis berharap karya ilmiah dapat diambil ilmu dan manfaatnya.

REFERENSI

- [1] Azis, A. B. A., & Rosli, R. (2021). Analisis Aras Kognitif Bagi Soalan Dalam Buku Teks Matematik Tahun 4 Kssr Semakan. *Malaysian Journal Of Social Sciences And Humanities (MJSSH)*, 6(3), 146-158.
- [2] Nathaniel, A., & Sannie, A. W. (2020). Analisis Semiotika Makna Kesendirian Pada Lirik Lagu “Ruang Sendiri” Karya Tulus. *SEMIOTIKA: Jurnal Ilmu Sastra Dan Linguistik*, 19(2), 107-117.
- [3] Widodo, B. E., & Purnomo, A. S. (2020). Implementasi Advanced Encryption Standard Pada Enkripsi Dan Dekripsi Dokumen Rahasia Ditintelkam Polda Diy. *Jurnal Teknik Informatika (Jutif)*, 1(2), 69-77.
- [4] Tanjung, R. Y., & Mesran, M. (2021). Perancangan Aplikasi Kompresi File Dokumen Menggunakan Algoritma Additive Code. *JURIKOM (Jurnal Riset Komputer)*, 8(4), 108-113.
- [5] Ahadia, N., Zulmuhibban, M., Aulia, C., Claudia, N., Apriyani, N., Widiawati, R., ... & Mertha, I. G. (2021). Pengenalan Dan Pengaplikasian Microsoft Word Dan Microsoft Powerpoint Di SMA Negeri 1 Praya Tengah. *Jurnal Pengabdian Magister Pendidikan IPA*, 4(1).
- [6] Sari, R., Fitriyani, A., & Prabandari, R. D. (2020). Optimalisasi Penggunaan Ms. Word Dan Ms. Excel Pada Siswa SMP PGRI Astra Insani Bekasi. *Jurnal Pengabdian Kepada Masyarakat UBJ*, 3(2).
- [7] Lindang, D. N., Muniar, A. Y., Halid, A., Muhajirin, M., & Amiruddin, A. (2023, February). Sistem Penentuan Kemiripan Antar Skripsi Menggunakan Metode Cosine Similarity Pada Perpustakaan. In *Seminar Nasional Teknik Elektro Dan Informatika (SNTEI)* (Vol. 8, No. 1, Pp. 321-324).
- [8] Habibi, M., & Cahyo, P. W. (2020). Journal Classification Based On Abstract Using Cosine Similarity And Support Vector Machine. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 4(3), 185-192.
- [9] Suryani, L., & Edy, K. (2020). PENGEMBANGAN APLIKASI "LOST & FOUND" BERBASIS ANDROID DENGAN MENGGUNAKAN METODE TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY. *Electro Luceat*, 6(2), 190-204.
- [10] Maulana, H. A. (2023). AUTOMATIC SUMMARIZING DOKUMEN REPOSITORY DENGAN TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY METHOD. *JURNAL TEKNIK INFORMATIKA (JUTEKIN)*, 11(1).