JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia

Abstrak—Kemudahan dalam mengakses informasi membawa berbagai keuntungan, termasuk kemampuan untuk mengembangkan model atau sistem yang dapat mendeteksi kemiripan antar dokumen, sistem pengecekan plagiarisme, pengelompokan teks berdasarkan tema, peringkasan otomatis, klasifikasi judul penelitian sesuai dengan topiknya, dan masih banyak lagi. Berbagai kegunaan ini membuat penelitian tentang deteksi kemiripan antar dokumen menjadi area yang penting untuk dikembangkan. Namun, studi mengenai deteksi kesamaan khusus untuk dokumen berbahasa Indonesia masih tergolong sedikit. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis komparatif terhadap kinerja Doc2Vec dibandingkan dengan Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance dalam mendeteksi kemiripan dokumen dengan teks berbahasa Indonesia. Tiga dataset digunakan dalam analisis ini, dengan dataset pertama terdiri dari 200 berita dari Google News, dataset kedua dari IndoNLU yang mempunyai 300 data, dan dataset ketiga dari TaPaCo dengan 1602 data. Temuan dari studi ini menunjukkan bahwa secara keseluruhan Cosine Similarity memiliki kinerja yang lebih baik dibandingkan Jaccard Coefficient dan Euclidean Distance dalam aspek akurasi, presisi, recall, dan nilai f-1. Di antara dataset yang diuji, Google News terbukti memiliki kinerja paling baik.

Kata Kunci: kemiripan teks Bahasa Indonesia, Doc2vec, Jaccard Coefficient, Cosine Similarity, Euclidean Distance

Abstract—*Ease of accessing information brings diverse benefits, including the ability to develop models that can detect similarities between documents, a plagiarism-checking system, automatic summarization, classification, etc. These various uses make research on similarity detection between documents a critical area to develop. However, studies about similarity detection specifically for Indonesian language documents are relatively few. Therefore, this research aims to conduct a comparative analysis of the performance of Doc2Vec compared to the Jaccard Coefficient, Cosine Similarity, and Euclidean Distance in detecting the similarity of documents with Indonesian text. Three datasets are used in this analysis, with the first dataset consisting of 200 news from Google News, the second dataset from IndoNLU, and the third dataset from TaPaCo. The findings from this study show that overall Cosine Similarity has better performance than Jaccard Coefficient and Euclidean Distance in all aspects. Among the tested datasets, Google News proved to perform the best.*

Keywords: similarity Indonesian text, Doc2vec, Jaccard Coefficient, Cosine Similarity, Euclidean Distance

1. PENDAHULUAN

Era pesatnya teknologi dan internet ini menyebabkan informasi dan data menyebar dengan sangat cepat dan mudah untuk didapatkan. Informasi dan data yang mudah didapatkan juga menyebabkan teknologi berkembang dengan pesat. Banyak manfaat dari teknologi yang pesat, salah satunya di bidang karya tulis. Hal positif dari pesatnya perkembangan teknologi adalah karya tulis dapat tersebar dengan luas dan pesat sehingga lebih populer dengan cepat. Penulis bisa mempromosikan karyanya dimana saja, kapan saja, dan tidak terbatas ruang dan waktu. Otomatisasi juga sudah mulai banyak dikembangkan untuk membantu pekerjaan manusia terutama dalam bidang karya tulis. Contohnya di bidang karya tulis adalah ada sistem yang mampu cek typo, sistem yang mampu cek grammar sebuah tulisan dalam bahasa asing, cek plagiat, cek kemiripan dua atau lebih dokumen, peringkasan paragraf otomatis, dan sebagainya. Misal saja dalam bidang pendidikan, seorang guru ingin koreksi jawaban esai dari muridnya. Namun guru tersebut juga harus cek satu-satu apakah ada esai yang sama atau saling contek atau tidak. Jika harus cek satu-satu, maka akan memakan waktu yang lama dan lebih sulit karena esai merupakan paragraf-paragraf yang panjang. Permasalahan tersebut bisa diatasi dengan cek kemiripan dokumen untuk mempermudah guru cek jawaban siswa tanpa memerlukan waktu yang lama. Namun, selain hal-hal positif, terdapat juga hal negatif yang diakibatkan dari cepatnya pergerakan data. Salah satunya adalah plagiasi karya, tulisan, atau sesuatu untuk kepentingan sendiri. Contohnya adalah seorang yang melakukan plagiasi, mahasiswa yang hanya copy-paste tanpa mencantumkan sumber, atau siswa yang mencontek karya atau tugas teman lainnya. Hal tersebut dapat diatasi dengan membuat sistem untuk cek kemiripan kata, kalimat, atau dokumen. Konsep ini mirip dengan cek plagiasi, satu dokumen diperiksa apakah dokumen tersebut mempunyai kesamaan yang sangat tinggi oleh dokumen-dokumen yang ada pada database. Jika sebuah dokumen mempunyai kesamaan kalimat atau paragraf pada dokumen-dokumen yang ada pada database, maka dapat dikatakan dokumen tersebut hanya copy-paste dari sebuah sumber. Sehingga, sistem ini bisa dikembangkan lebih lanjut menjadi cek plagiasi, tidak hanya cek kesamaan dokumen saja.

Beberapa penelitian terdahulu yang telah dilakukan tentang cek kemiripan kata adalah metode similarity untuk kemampuan maksimum interpretasi manusia pada 2019 oleh Sitikhu dkk. Metode-metode yang dilakukannya antara lain Cosine Similarity dengan TF-IDF, Cosine Similarity dengan Word2Vec Vectors, dan Soft Cosine Similarity dengan Word2Vec Vectors [1]. Penelitian cek kemiripan kata juga dilakukan oleh beberapa peneliti untuk kemiripan antara teks atau dokumen serta objek yang bukan teks juga menggunakan metode jarak dan kemiripan dua objek seperti Cosine Similarity [2], [3], [4], Jaccard Coefficient[5], gabungan dari kedua metode tersebut [6], dan Euclidean Distance [7]. Selanjutnya peneliti pada tahun 2022, Hendrawan dkk. melakukan sebuah penelitian komparasi word2vec dan doc2vec. Penelitian tersebut menyatakan bahwa Doc2vec mempunyai kelebihan dari sisi mengukur kedekatan per kalimat. Hasil penelitian menunjukkan performa Doc2Vec lebih unggul [8]. Penggunaan Doc2Vec juga diteliti oleh beberapa penelitian seperti [9], [10], [11]. Pada penelitian [12]

Commented [A1]: Penulisan Judul Artikel cukup bagus. Pada judul sudah memiliki masalah yang di bahas, metode/solusi penyelesaian masalah, dan informatif. Sebaiknya Judul memiliki kata sebanyak 14-18 kata.

Commented [A2]: Penulisan Abstrak cukup bagus. Pada Abstrak sudah memiliki masalah yang di bahas pada penelitian, solusi/metode yang digunakan, tujuan dan kontribusi dari penelitian, serta belum jelas hasil sementara yang dicapai.

Commented [A3]: Penulisan Isi pendahuluan cukup bagus. Isi pendahuluan sudah menggambarkan masalah penelitian, metode pembanding, penelitian sejenis/terkait, tujuan penelitian yang akan dilakukan, namun belum menguraikan secara jelas GAP/Perbedaan dari penelitian sebelumnya dengan penelitian yang dilakukan, sudah mengkaitkan teori yang digunakan dengan rujukan/ referens yang terdapat pada DAFTAR PUSTAKA, serta memiliki pernyataan kontribusi dari hasil penelitian. Referensi/kutipan sudah ditulis dengan format IEEE yang menggunakan Soft Referensi Ilmiah.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

juga menggunakan Word2Vec yang kemudian menyarankan untuk menggunakan Doc2Vec. Selain itu, Doc2Vec juga bisa mengubah teks menjadi suara seperti pada penelitian [13], [14].

Penelitian terdahulu seperti penelitian oleh Sitikhu dan Singh menyarankan cek kemiripan antar dokumen dengan menggunakan Doc2Vec agar akurasi menjadi lebih baik. Oleh karena itu, penelitian ini melakukan komparasi metode cek kemiripan dokumen Bahasa Indonesia menggunakan metode Doc2Vec. Metode similarity yang digunakan adalah metode Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Beberapa menggunakan algoritma ini sebagai metode untuk cek kemiripan dikarenakan metode tersebut pas untuk menghitung kemiripan dua objek. Penelitian terdahulu mengenai metode ini salah satunya adalah komparasi Cosine Similarity, Jaccard Coefficient, dan Euclidean Distance pada teks bahasa Arab oleh Alobed, dkk. pada tahun 2021 [15]. Selain itu, ada penelitian mengenai ringkasan teks dengan metode similarity menggunakan Cosine Similarity [16], Jaccard Coefficient, dan Euclidean Distance[17]. Tujuan dari penelitian ini adalah komparasi performa metode-metode tersebut untuk dokumen yang menggunakan Bahasa Indonesia.

Banyak manfaat dari cek kemiripan teks dan terdapat beberapa metode untuk cek kemiripan teks dokumen dengan berbagai bahasa. Banyak pula penelitian mengenai hal ini dengan berbagai bahasa. Hal ini menunjukkan kemiripan teks banyak digunakan dan dimanfaatkan di seluruh penjuru dunia. Namun penelitian untuk teks yang menggunakan bahasa Indonesia masih sedikit jumlahnya. Oleh karena itu, pengembangan lebih lanjut mengenai cek kemiripan teks berbahasa Indonesia perlu ditingkatkan. Kontribusi penelitian ini untuk menunjang pengembangan pengolahan teks bahasa Indonesia adalah dengan dilakukannya analisis komparasi berbagai metode untuk cek kemiripan dengan teks Bahasa Indonesia. Namun, tidak semua kasus dapat dilakukan pada penelitian ini, sehingga pada penelitian ini, terdapat beberapa batasan masalah. Batasan masalah penelitian ini yaitu penelitian ini menggunakan dua dataset yang dapat diakses secara bebas yaitu IndoNLU dan TaPaCo dan mengumpulkan satu dataset berita secara manual yaitu Google News. Google News berisi 200 berita Bahasa Indonesia diambil dari bulan Juni sampai Agustus 2022 secara acak dan berita tersebut telah dikelompokkan oleh Google News. Penelitian ini juga dibatasi oleh dataset yang menggunakan Bahasa Indonesia saja.

2. METODOLOGI PENELITIAN

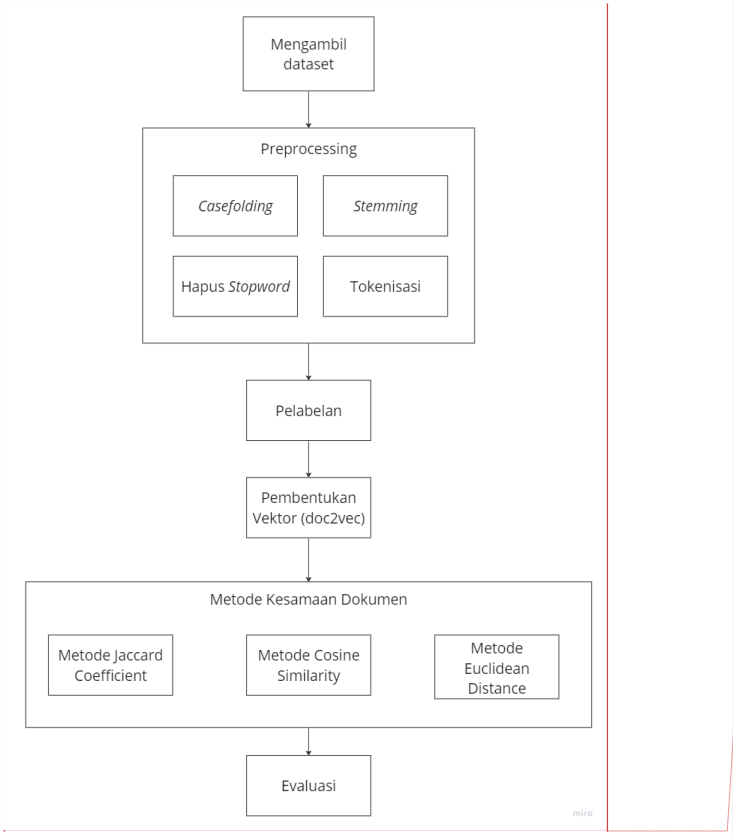
Commented [A4]: Penulisan metodologi penelitian cukup bagus. Bagian metodologi ini sudah memiliki tahapan penelitian yang menggambarkan tahapan apa yang dilakukan pada penelitian, terlihat penerapan solusi/metode pada tahapan penelitian, serta memiliki kajian pustaka dari algoritma/metode yang digunakan. Setiap penulisan sudah memiliki referensi/kutipan dengan format IEEE yang ditulis menggunakan Soft Referensi Ilmiah.

2.1 Tahapan Penelitian

Dua buah objek yang dikatakan mirip adalah jika dua objek tersebut hampir sama atau serupa. Misal sebuah berita dari website A membicarakan mengenai kenaikan harga minyak di Indonesia, sedangkan berita dari website B juga demikian namun isi dalam berita tersebut tidak sama persis, maka kedua berita tersebut tetap dikatakan mirip karena membahas persoalan yang sama. Hal tersebut terjadi karena jika kedua berita tersebut sama persis padahal dari sumber/media yang berbeda dan tidak menyatakan sumber asli berita tersebut, bisa terkena plagiasi. Penelitian ini merupakan komparasi kemiripan atau similarity untuk teks atau dokumen berbahasa Indonesia. Metode yang digunakan untuk cek similaritas adalah Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Sedangkan untuk metode pembuatan vektor dokumen yang digunakan adalah Doc2Vec. Langkah awal dalam penelitian ini adalah mengumpulkan dataset. Kemudian dataset dimasukkan ke tahap proses preprocessing. Tahap preprocessing ada empat proses, yaitu case folding, stemming, penghapusan stopword, dan pembuatan token (Tokenisasi). Setelah preprocessing data, dataset tersebut dibentuk vektor dokumen dengan metode yang telah disebutkan. Langkah akhir adalah perhitungan kemiripan antara dua dokumen yang diikuti dengan evaluasi. Tahap evaluasi dilakukan dengan menganalisis hasil. Skema rancangan umum penelitian ini ditunjukkan pada Gambar 1.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293



Gambar 1. Tahapan Penelitian Secara Umum

2.2 *Preprocessing*

Preprocessing data yang dilakukan pada penelitian ini ada empat, yaitu case folding, stemming, menghapus stopwords, dan pembuatan token. Case folding membuat semua huruf menjadi huruf kecil. Setelah itu, proses menghapus kata imbuhan atau kata yang tidak perlu dan diikuti dengan penghapusan kata depan. Selanjutnya melakukan tokenisasi. Contoh hasil preprocessing dilihat pada Tabel 1 sebagai berikut.

Tabel 1. Contoh *Preprocessing*

Proses	**Hasil *Preprocessing***
Awal	Krjogja.com - YOGYA - Badan Meteorologi Klimatologi dan Geofisika (BMKG) DIY kembali mengeluarkan peringatan dini terhadap gelombang laut di perairan Yogyakarta, Senin (04/03/2024).
Casefolding	krjogja.com - yogya - badan meteorologi klimatologi dan geofisika (bmkg) diy kembali mengeluarkan peringatan dini terhadap gelombang laut di perairan yogyakarta, senin (04/03/2024).
Stemming	krjogjacom yogya badan meteorologi klimatologi dan geofisika bmkg diy kembali keluar peringatan dini hadap gelombang laut di air yogyakarta senin 04032024
Hapus *Stopwords*	krjogjacom yogya badan meteorologi klimatologi geofisika bmkg diy kembali keluar peringatan dini hadap gelombang laut air yogyakarta senin 04032024
Tokenisasi	“krjogjacom”,“yogya”,“badan”,“meteorologi”,“klimatologi”,“geofisika”,“bmkg”,“diy ”,“kembali”,“keluar”,“peringatan”,“dini”,“hadap”,“gelombang”,“laut”,“air”, “yogyakarta”,“senin”,“04032024”

Commented [A5]: Gambar diberikan penomoran dan judul min 2 kata, Gambar terlihat jelas, tidak berwarna dan berkualitas baik. Setiap Gambar diberikan penjelasan detail dan mengkaitkan penomoran gambar pada isi penjelasan yang dilakukan. Sebelum gambar harus di berikan kalimat pengantar.

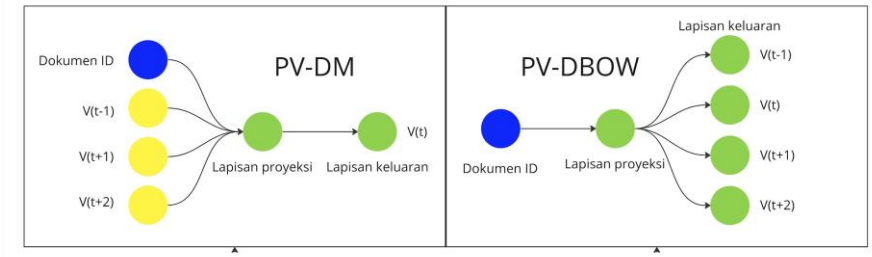
Commented [A6]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

2.3 Doc2Vec

Doc2vec dikembangkan oleh pengembang Word2Vec juga, yaitu Le dan Mikolov pada tahun 2014 yang digunakan untuk membuat representasi numerical dari dokumen (vektor) dengan mempertimbangkan konteks atau kata-kata yang muncul [18]. Doc2vec termasuk algoritma unsupervised. Doc2vec dan word2vec mempunyai satu perbedaan, yaitu word2vec menggunakan kata (word) sedangkan doc2vec menggunakan dokumen atau paragraf untuk direpresentasikan ke dalam vektor. Doc2Vec menggunakan dua algoritma utama yaitu Paragraph Vector - Distributed Memory (PV-DM) dan Paragraph Vector - Distributed Bag of Words (PV-DBOW). Cara kerja distributed memory adalah dengan menggunakan dua input, yaitu ID dokumen dan kata-kata yang ada pada dokumen. Setelah itu akan masuk pada lapisan proyeksi yang bertugas membuat vektor kata dan vektor dokumen. Vektor-vektor tersebut masuk ke dalam fungsi yang meminimalkan perbedaan antara kada prediksi dan kata target. Setelah itu, output akan terlihat pada lapisan keluaran. Kebalikan dari DM, PV-DBOW berfokus pada kat-kata yang distribusikan dan bukan maknanya. Sehingga algoritma ini pas untuk melihat pola teks, bukan isi sebuah teks. PV-DBOW hanya mempunyai satu input, yaitu ID dokumen. ID tersebut yang nantinya menjadi vektor dokumen. Vektor yang dihasilkan dari doc2vec ini bisa digunakan untuk menghitung kemiripan antar dokumen. Berikut arsitektur Doc2Vec ditampilkan pada Gambar 2.



Gambar 2. Arsitektur Doc2Vec

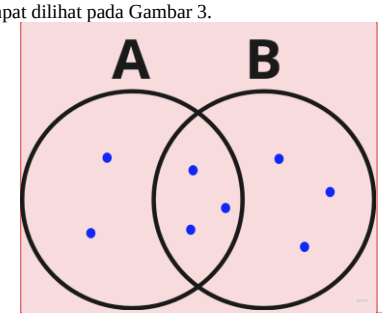
Commented [A7]: Gambar diberikan penomoran dan judul min 2 kata, Gambar terlihat jelas, tidak berwarna dan berkualitas baik. Setiap Gambar diberikan penjelasan detail dan mengkaitkan penomoran gambar pada isi penjelasan yang dilakukan. Sebelum gambar harus di berikan kalimat pengantar.

2.4 Jaccard Coefficient

Banyak metode yang bisa digunakan untuk cek kemiripan kata atau kalimat, salah satunya dengan Jaccard Coefficient atau Jaccard Index. Jaccard Coefficient membutuhkan dua dokumen untuk dicari kemiripannya [19]. Jaccard Coefficient dari dua himpunan A dan B dapat dilihat dari persamaan nomor 1.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Sehingga dapat disimpulkan Jaccard Coefficient dari himpunan A dan B adalah jumlah elemen (kardinal) dari irisan himpunan A dan himpunan B kemudian dibagi dengan kardinal dari gabungan himpunan A dan himpunan B. Sebagai contoh dapat dilihat pada Gambar 3.



Gambar 3. Ilustrasi Jaccard Coefficient Himpunan A dan B

Commented [A8]: Gambar diberikan penomoran dan judul min 2 kata, Gambar terlihat jelas, tidak berwarna dan berkualitas baik. Setiap Gambar diberikan penjelasan detail dan mengkaitkan penomoran gambar pada isi penjelasan yang dilakukan. Sebelum gambar harus di berikan kalimat pengantar.

Gambar 3 menunjukkan kardinal dari himpunan A adalah 5, sedangkan kardinal himpunan B adalah 6. Kardinal dari irisan himpunan A dan B adalah 3. Kardinal dari gabungan himpunan A dan B adalah 8 item. Sehingga dapat disimpulkan untuk Jaccard Coefficient dari himpunan A dan B adalah 3/8. Contoh perhitungan Jaccard dapat dilihat sebagai berikut.

Terdapat 2 vektor (vektor A dan vektor B).

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

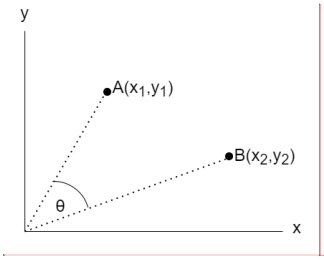
$$A = [1, 3, 6, 9], B = [1, 2, 3, 4]$$

Sehingga dengan persamaan nomor 1, Jaccard Coefficient dapat dari vektor A dan B adalah

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{2}{6} = 0.33$$

2.5 Cosine Similarity

Cosine Similarity adalah salah satu metode menghitung kemiripan antara dua kata/kalimat. Sebelum dihitung kemiripannya, kata/kalimat tersebut diubah dalam bentuk vektor. Setelah itu, Cosine Similarity akan menghitung kemiripan dua vektor dengan menghitung nilai cosinus dari sudut terkecil dari vektor tersebut. Cosine Similarity terletak antara -1 dan 1. Semakin mendekati 1 maka semakin mirip vektor tersebut [20]. Konsep cosine similarity dapat dilihat pada Gambar 4.



Commented [A9]: Gambar diberikan penomoran dan judul min 2 kata, Gambar terlihat jelas, tidak berwarna dan berkualitas baik. Setiap Gambar diberikan penjelasan detail dan mengkaitkan penomoran gambar pada isi penjelasan yang dilakukan. Sebelum gambar harus di berikan kalimat pengantar.

Gambar 4. Ilustrasi Cosine Similarity Titik A dan B

Diketahui dua vektor A dan B dengan sudut theta. Similaritas pada kedua vektor tersebut adalah dengan menggunakan persamaan nomor 2.

$$sim(A,B) = cos\ \theta = \frac{A \cdot B}{|A||B|} = \frac{x_1x_2 + y_1y_2}{\sqrt{(x_1{}^2 + y_1{}^2)(x_2{}^2 + y_2{}^2)}} \quad (2)$$

Dimana A·B adalah dot product dari vektor A dan B. Dot product bisa dihitung dengan cara mengalikan vektor A dengan vektor B. |A| adalah panjang dari vektor A dan |B| adalah panjang dari vektor B. Contoh untuk perhitungan cosine adalah sebagai berikut. Terdapat 2 vektor (vektor A dan vektor B).

$$A = [2, 3, 4, 1], B = [1, 2, 3, 4]$$

Maka Cosine Similarity dari vektor A dan B adalah

$$Cos(A,B) = \frac{\underline{A} \cdot \underline{B}}{|\underline{A}| \cdot |\underline{B}|}$$

$$= \frac{(2 \times 1) + (3 \times 2) + (4 \times 3) + (1 \times 4)}{\sqrt{2^2 + 3^2 + 4^2 + 1^2} \cdot \sqrt{1^2 + 2^2 + 3^2 + 4^2}}$$

$$= \frac{2 + 6 + 12 + 4}{\sqrt{4 + 9 + 16 + 1} \cdot \sqrt{1 + 4 + 9 + 16}}$$

$$= \frac{24}{\sqrt{30} \cdot \sqrt{30}} = \frac{24}{30} = 0.8$$

2.6 Euclidean Distance

Euclidean Distance adalah metode untuk menghitung jarak. Semakin kecil nilai jarak, semakin mirip antara satu vektor dengan vektor lainnya. Rumus yang digunakan untuk menghitung euclidean distance dapat dilihat pada persamaan nomor 3 [21].

$$d = \sqrt{\sum_{i=1}^{n} (pi - qi)^2} \quad (3)$$

Jarak dari vektor p ke vektor q adalah d dan n adalah jumlah elemen dari vektor p dna q. Sedangkan pi dan qi masing-masing adalah elemen ke-1 dari vektor p dan q.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

3. HASIL DAN PEMBAHASAN

Commented [A10]: Penulisan hasil dan pembahasan cukup bagus. Isi Bagian ini sudah menguraikan tahapan dari penerapan algoritma/metode dalam menyelesaikan masalah, serta hasil yang di peroleh dari algoritma/metode yang digunakan sudah ada.

Implementasi menggunakan python 3 dengan berbagai library. Pada proses implementasi, penelitian ini menggunakan parameter yang telah ditentukan. Parameter tersebut dapat dilihat pada tabel 2 sebagai berikut.

Tabel 2. Tabel Parameter

Parameter	Nilai
DM	1, 0
Max Epoch	10, 20, 100, 400
Vector Size	10, 20, 100
Window	5, 15
Min Count	1

Commented [A11]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

DM adalah jenis doc2vec yang digunakan, 1 adalah Distributed Memory dan 0 adalah Distributed Bag of Words. Sedangkan max epoch adalah maksimum perulangan pada model agar lebih akurat. Kemudian vec size adalah ukuran vektor. Window adalah jarak antara kata saat ini dengan kata yang menjadi input, dan min count adalah minimal kata yang muncul. Implementasi diawali dengan mendapatkan dataset. Penelitian ini menggunakan tiga dataset. Dataset pertama adalah kumpulan berita dengan total 200 berita yang isinya diambil secara manual dari Google News pada tautan pada tautan https://news.google.com/home?hl=id&gl=ID&ceid=ID:id. Berita yang digunakan ada empat topik yang berbeda. Dataset kedua adalah dataset dari IndoNLU yang berisikan kumpulan data bisa berupa kalimat atau paragraf berbahasa Indonesia. IndoNLU bisa diakses di https://github.com/IndoNLP/indonlu. Dataset ketiga adalah dataset dari TaPaCo. TaPaCo adalah dataset yang terdiri dari kalimat dan parafrasenya dan yang diambil dari dataset TaPaCo adalah kalimat yang menggunakan bahasa Indonesia saja. Dataset TaPaCo bisa diakses di https://huggingface.co/datasets/tapaco.

Setelah mendapatkan dataset, preprocessing dilakukan untuk membersihkan data dari kata-kata yang tidak perlu agar proses menjadi lebih cepat. Setelah itu, pembentukan model. Model yang terbentuk menjadi tiga, yaitu Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Setelah model didapatkan, selanjutnya proses evaluasi. Hasil diperoleh dari melakukan tahapan evaluasi dengan membandingkan hasil label asli dengan hasil label yang diperoleh dari model. Hasil berupa akurasi, recall, precision, f1-score, dan waktu training yang didapatkan. Perhitungan evaluasi tersebut menggunakan Confusion Matrix.

3.1 Performa Model

Akurasi menandakan nilai dokumen benar diklasifikasikan dengan keseluruhan dokumen yang menjawab persoalan perbandingan dokumen mirip dan tidak mirip diprediksi benar dengan keseluruhan dokumen. Precision dan recall yang digunakan adalah rata-rata setiap kelas. Kelas pada penelitian ini ada dua, kelas 1 yang berarti kedua dokumen tersebut memiliki nilai kemiripan yang tinggi, sedangkan kelas 0 yang berarti kedua dokumen tersebut memiliki kemiripan yang rendah. Precision kelas 1 dan kelas 0 dihitung prediksi benar untuk kelas tersebut dibandingkan dengan keseluruhan data yang masuk ke dalam kelas tersebut. Setelah itu, dicari rata-rata precision untuk kelas 1 dan kelas 0. Precision untuk masing-masing kelas menjawab persoalan perbandingan berita yang terklasifikasi dengan benar dengan keseluruhan dokumen yang terklasifikasi tersebut. Kemudian recall adalah perbandingan yang terklasifikasi benar dibandingkan dengan keseluruhan yang sesungguhnya. Sedangkan, f-1 score adalah perbandingan rata-rata precision dan recall.

Hasil keseluruhan performa, didapatkan performa terbaik menggunakan Cosine Similarity karena memiliki rata-rata performa yang lebih baik daripada Jaccard Coefficient dan Euclidean Distance. Hal ini dikarenakan perhitungan Jaccard yang sederhana serta perhitungan Euclidean Distance yang kurang pas untuk kasus ini. Jaccard Coefficient membandingkan angka-angka pada vektor-vektor kemudian menghitung berapa angka yang sama dan yang berapa angka yang tidak sama. Contoh kasus dua dokumen sebenarnya mirip, namun tidak sama persis. Hal ini menyebabkan vektor juga tidak sama persis dan dapat menyebabkan kurangnya nilai kesamaan antara dua dokumen tersebut. Jika ingin performanya lebih baik, maka model harus lebih baik agar vektor yang dihasilkan memiliki vektor yang mempunyai anggota yang nilainya sama semua. Sedangkan jika menggunakan Euclidean Distance, nilai kemiripan yang dihasilkan berkisar antara 0 sampai tak hingga. Hal tersebut membuat perhitungan harus dijadikan sedemikian rupa atau dinormalisasi sehingga 0 sampai 1 saja, dengan 0 berarti tidak ada kemiripan antara dokumen-dokumen tersebut. Namun, batas-batas kemiripan lebih jelas menggunakan Cosine Similarity dan Jaccard Coefficient.

Kelebihan Cosine Similarity dibandingkan metode yang lain adalah Cosine Similarity menghitung sudut dari vektor-vektor tersebut. Kedua vektor yang dibandingkan lalu dilihat sudutnya dan dihitung nilai cosine. Jika kedua vektor tersebut didekatkan dan mempunyai sudut yang mendekati 0º, maka nilai cosine akan bernilai 1 yang menandakan bahwa kedua vektor tersebut sangat dekat dan dapat dikatakan kedua vektor atau dokumen tersebut mirip. Begitu juga sebaliknya, jika kedua vektor tersebut mendekati 90º, maka nilai cosine adalah 0. Sehingga dapat dikatakan kedua vektor tersebut tidak mirip. Vektor-vektor dalam kasus ini bisa lebih akurat dalam klasifikasinya dengan menggunakan Cosine Similarity karena tidak harus ada unsur yang sama, namun bisa dikatakan dokumen atau vektor yang merepresentasikan dokumen tersebut berdekatan, maka akan dinilai kedua

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

dokumen tersebut mirip. Sebagai contoh untuk memperjelas perbedaan cara perhitungan Cosine Similarity, Jaccard Coefficient, dan Euclidean Distance, pada salah satu hasil percobaan pada dataset IndoNLU, pada sentence A dan sentence B terdapat kalimat berikut.

"Salah satu tekniknya adalah periplus , deskripsi pada pelabuhan dan daratan sepanjang garis pantai yang bisa dilihat pelaut di lepas pantai ; contoh pertamanya adalah Hanno sang Navigator dari Carthaginia dan satu lagi dari Laut Erythraea , keduanya selamat di laut menggunakan teknik periplus dengan mengenali garis pantai laut Merah dan Teluk Persi"

"Bangsa Romawi memberi sumbangan pada pemetaan karena mereka banyak menjelajahi negeri dan menambahkan teknik baru"

Kedua dokumen tersebut berlabel 0, yang artinya tidak mirip. Kedua dokumen tersebut diproses ke preprocessing dan doc2vec. Setelah perhitungan doc2vec untuk mendapatkan nilai vektor dari kedua dokumen tersebut, didapat dua vektor [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0] dan [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]. Setelah itu, vektor tersebut dihitung dengan Cosine Similarity. Perhitungan Cosine Similarity pada kasus ini adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Cosine(A,B) = \frac{(2\times 0) + \big((-6)\times 1\big) + (1\times 0) + \cdots + (6\times 0) + ((-1)\times 0) + ((-2)\times(-1))}{\sqrt{2^2 + (-6)^2 + 1^2 + \cdots + 6^2 + (-1)^2 + (-2)^2} \cdot \sqrt{0^2 + 1^2 + 0^2 + \cdots + 0^2 + 0^2 + (-1)^2}}$$

$$= \frac{0 + (-6) + 0 + 8 + 0 + 3 + 0 + 0 + 0 + 2}{\sqrt{4 + 36 + 1 + 16 + 1 + 9 + 0 + 36 + 1 + 4} \cdot \sqrt{0 + 1 + 0 + 4 + 0 + 1 + 1 + 0 + 0 + 1}}$$

$$= \frac{7}{\sqrt{108} \cdot \sqrt{8}} \approx \frac{7}{10.39 \cdot 2.83} \approx 0.24$$

Sehingga kedua dokumen ini memiliki nilai kemiripan 0.24 yang berarti tidak mirip. Mirip tidaknya menggunakan pembulatan biasa, yaitu 0.5 dikatakan kedua dokumen tersebut mirip. Namun jika 0 nilai kemiripan <0.5 , maka kedua dokumen tersebut tidak mirip. Sedangkan untuk perhitungan Jaccard Coefficient adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Jaccard(A,B) = \frac{n(\{-1.0, 0.0, 1.0, 2.0\})}{n(\{-6.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0, 4.0, 6.0\})} = \frac{4}{9} \approx 0.44$$

Jaccard Coefficient menghasilkan nilai 0.44 yang berarti tidak mirip juga walau nilai kemiripan berbeda dan Jaccard hampir mendekati nilai 0.5. Jaccard memiliki keuntungan berupa perhitungannya yang cepat karena tidak rumit. Hanya membandingkan anggota vektor mana saja yang sama dengan keseluruhan anggota vektor. Namun Jaccard juga memiliki kekurangan, perhitungannya yang sangat simple tersebut menyebabkan hitungan tidak akurat, Hanya karena berbeda beberapa anggota vektor, bukan berarti vektor tersebut tidak dekat, begitu juga sebaliknya. Sedangkan perhitungan dengan Euclidean Distance adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Euclidean\ (A,B) = \sqrt{(2-0)^2 + ((-6)-1)^2 + \cdots + ((-1)-0)^2 + ((-2)-(-1))^2}$$

$$Euclidean\ (A,B) = \sqrt{4 + 49 + 1 + 4 + 1 + 4 + 1 + 36 + 1 + 1} = \sqrt{102} \approx 10.1$$

Hasil menunjukkan dengan Euclidean Distance jarak yang diperoleh adalah 10.1 dan setelah melalui proses normalisasi dengan nilai dari max dari Euclidean, nilai kemiripan yang diperoleh adalah 0.64. Apabila nilai dibulatkan, nilai mendekati 1. Artinya dengan Euclidean Distance, kedua dokumen tersebut bernilai mirip. Kelebihan dari Euclidean Distance ini perhitungannya tidak lebih rumit daripada Cosine Similarity. Namun, pada kasus ini untuk mencari kemiripan dari dua vektor tidak cukup hanya mengandalkan jarak dari vektor-vektor tersebut. Pola dari vektor serta kemiripan vektor itu sendiri perlu juga untuk diperhatikan. Semua parameter telah dicoba dan terdapat total 48 percobaan dari parameter yang dicoba satu-satu. Rata-rata hasil performa untuk semua dataset ditampilkan pada Tabel 3.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Tabel 3. Tabel Perbandingan Rata-rata Performa Tiga Metode

Performa	Jaccard Coefficient	Cosine Similarity	Euclidean Distance
Akurasi	0.39	0.69	0.57
Precision	0.56	0.58	0.59
Recall	0.55	0.73	0.48
F1-Score	0.29	0.49	0.47
Time (s)	8.52	23.82	20.54

Commented [A12]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

Dari hasil tersebut, dapat dilihat jika rata-rata keseluruhan untuk performa Cosine lebih baik dibandingkan dengan Jaccard dan Euclidean, akan tetapi waktu proses perhitungan yang dilakukan Cosine lebih besar daripada algoritma lainnya. Hal ini dipengaruhi oleh perhitungan Cosine yang lebih kompleks dibandingkan dengan algoritma lainnya. Rata-rata tersebut berdasarkan 48 percobaan dari berbagai parameter, kemudian dicari percobaan mana saja yang pas untuk tiap-tiap metode.

3.2 Analisis Dataset

Dataset Google News yang diambil secara manual terdiri dari 200 berita dan mencakup empat tema dengan masing-masing tema diwakili oleh 40 berita. Dataset ini mempunyai jumlah total kata sebanyak 71.085 kata. Rata-rata kata dalam satu berita adalah sekitar 355 kata. Sementara itu, Dataset IndoNLU yang terdiri dari 300 dokumen atau paragraf, memiliki total kata sebanyak 6.767, dengan rata-rata jumlah kata per dokumen adalah 22 kata. Dataset TaPaCo, yang meliputi 1.602 dokumen atau paragraf, memiliki total 7.889 kata dengan rata-rata jumlah kata per dokumen berkisar antara 4 sampai 5 kata saja. Berdasarkan analisis jumlah dan keragaman kata, Dataset Google News memiliki keragaman kata dan volume yang paling tinggi dibandingkan dengan Dataset IndoNLU dan TaPaCo, yang berada pada urutan berikutnya.

Performa Dataset Google News menunjukkan hasil yang superior dibandingkan dengan dua dataset lainnya, dengan mencapai akurasi sebesar 0,98, presisi 0,84, nilai recall 0,95, dan skor F1 0,89 dalam waktu pelatihan 10,56 detik. Di sisi lain, Dataset IndoNLU menunjukkan kinerja maksimal dengan akurasi 0,72, presisi 0,75, nilai recall 0,65, dan skor F1 0,65 dalam waktu pelatihan yang sangat singkat, yaitu 0,003 detik. Dataset TaPaCo, dengan kinerja terbaiknya, menunjukkan akurasi tertinggi sebesar 0,99, presisi 0,63, nilai recall 0,89, dan skor F1 0,69 dalam waktu pelatihan 24,49 detik. Temuan ini mengindikasikan bahwa model doc2vec yang diterapkan dalam penelitian ini cenderung menghasilkan kinerja yang lebih baik pada dataset dengan keragaman kata yang lebih luas namun tetap terfokus pada satu topik, di mana satu dokumen mengandung paragraf yang terdiri dari jumlah kata yang lebih banyak.

Hal tersebut juga dipengaruhi oleh dokumen pada dataset lain mempunyai padanan kata yang sedikit. Misal pada dataset TaPaCo ada dua dokumen yang berisi kalimat “Anak perempuan mirip dengan ibunya.” dan “Buah apel tidak jatuh jauh dari pohonnya.”. Secara konteks, dua dokumen atau kalimat ini sama. Namun secara kemiripan, dua dokumen tersebut dapat dikatakan tidak mirip karena mempunyai struktur kalimat yang berbeda. Begitu juga dengan dataset IndoNLU. Berbeda dengan Google News. Google News berisi berita-berita yang tentunya jika ada sebuah berita, maka media satu dengan yang lainnya akan mirip saat menyampaikan berita tersebut secara tertulis. Jika berita yang berbeda, maka konteks dari kedua berita tersebut berbeda pula. Sehingga untuk keseluruhan performa, Google News lebih baik daripada dataset lainnya. Sedangkan waktu training, dipengaruhi oleh banyaknya dokumen dan banyaknya kata dalam suatu dokumen. Untuk informasi detail mengenai dataset dan kinerjanya dengan menggunakan model doc2vec, dapat dirujuk pada Tabel 4.

Tabel 4. Tabel Performa Terbaik Dataset

Performa	Google News	IndoNLU	TaPaCo
Banyak dokumen	200	300	1602
Total kata	71085	6767	7889
Rata-rata kata	355	22	4-5
Akurasi	0.98	0.72	0.99
Presisi	0.84	0.75	0.63
Recall	0.95	0.65	0.89
f-1 score	0.89	0.65	0.69
Waktu *training* (detik)	10.56	0.003	24.49

Commented [A13]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

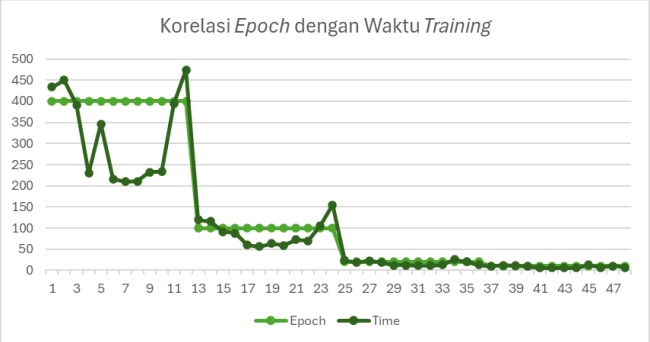
3.3 Analisis Parameter

Hasil kinerja terbaik untuk parameter setiap metode kemiripan secara umum dicapai dengan metode doc2vec 0, yang berarti menggunakan metode Distributed Bag of Words. Distributed Bag of Words terbaik diperoleh dari metode kemiripan Cosine Similarity dan Euclidean Distance. Sedangkan performa terbaik Jaccard Coefficient menggunakan doc2vec Distributed Memory. Maksimal epoch terbaik adalah 400 pada Jaccard dan Euclidean, sedangkan maksimal epoch terbaik 10 pada Cosine. Epoch menandakan banyaknya perulangan yang terjadi pada model agar memperoleh hasil yang terbaik. Cosine Similarity termasuk singkat dengan 10 epoch.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Sedangkan Jaccard dan Euclidean membutuhkan 400 epoch agar hasil maksimal. Hal ini juga berpengaruh pada waktu yang dibutuhkan sebuah model untuk menjalankan tugasnya. Semakin besar epoch, semakin besar pula waktu yang dibutuhkan dalam satu kali training. Walau waktu training lebih lama daripada percobaan lainnya, namun pada Jaccard dan Euclidean menghasilkan performa yang paling baik. Korelasi epoch dengan waktu training diilustrasikan pada Gambar 5.



Gambar 5. Korelasi *Epoch* dan Waktu Training

Dengan demikian, pendekatan menggunakan Cosine dengan 10 epoch dinilai lebih efektif karena membutuhkan waktu yang singkat dengan performa terbaik. Ukuran vektor yang menghasilkan kinerja terbaik untuk semua dataset adalah 100 untuk Cosine dan 10 untuk Jaccard dan Euclidean. Ukuran vektor yang dimaksud adalah ukuran pada suatu vektor, apabila 10 maka ada 10 elemen pada satu vektor. Selain itu, ukuran window yang optimal dalam penelitian ini adalah 5 untuk Cosine dan Euclidean dan 10 untuk Jaccard. Window menunjukkan banyaknya kata yang terlibat. Parameter terbaik yang diterapkan dalam penelitian ini dilihat pada Tabel 5.

Tabel 5. Parameter Terbaik

Similarity	Doc2Vec	Epoch	Vektor	Window
Jaccard	Distributed Memory	400	10	15
Cosine	Distributed Bag of Words	10	100	5
Euclidean	Distributed Bag of Words	400	10	5

3.4 Analisis Performa Dataset Google News

Implementasi pada dataset Google News menunjukkan bahwa Cosine Similarity dengan penggunaan Distributed Memory (DM), epoch maksimal 10, ukuran vektor 100, dan window 5, menghasilkan model dengan performa terbaik untuk dataset ini. Waktu training model ini selama 10.56 detik dan waktu proses Cosine selama 1.6 detik, dengan hasil akurasi sebesar 0.98, presisi 0.84, recall 0.95, dan nilai F-1 0.89. Di sisi lain, metode Jaccard memerlukan waktu pelatihan yang lebih lama yaitu 433.61 detik dengan akurasi terbaik yang dicapai pada epoch 400, menggunakan Distributed Memory (DM), ukuran vektor 20, dan window 15, menghasilkan akurasi 0.98. Hal ini menunjukkan bahwa meskipun Jaccard mencapai akurasi yang hampir sama dengan Cosine, namun waktu yang diperlukan jauh lebih lama, dengan nilai recall dan F-1 yang tidak sebaik Cosine, sehingga kinerjanya dianggap kurang optimal. Pendekatan menggunakan metode Euclidean dengan doc2vec Distributed Memory (DM), epoch 400, ukuran vektor 10, dan window 5 mendapatkan akurasi 0.8, presisi 0.48, recall 0.42, dan nilai F-1 0.45, serta waktu proses Euclidean terlama dibanding metode yang lain yaitu 17.24 detik dan dan waktu pelatihan 391.31 detik. Berdasarkan analisis ini, Cosine Similarity terbukti sebagai metode terbaik untuk dataset ini berkat efisiensi waktu dan kinerja yang unggul. Dataset ini juga lebih pas menggunakan doc2vec Distributed Memory (DM) dibandingkan dengan Distributed Bag of Words (DBOW). Ringkasan hasil implementasi pada dataset Google News tersaji dalam Tabel 6.

Tabel 6. Ringkasan Hasil Dataset Google News

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	1	10	100	5	0.98	0.84	0.95	0.89	10.56
Cosine	1	400	20	15	0.98	0.99	0.67	0.75	433.61
Euclidean	1	400	10	5	0.80	0.48	0.42	0.45	391.31

Commented [A14]: Gambar diberikan penomoran dan judul min 2 kata, Gambar terlihat jelas, tidak berwarna dan berkualitas baik. Setiap Gambar diberikan penjelasan detail dan mengkaitkan penomoran gambar pada isi penjelasan yang dilakukan. Sebelum gambar harus di berikan kalimat pengantar.

Commented [A15]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

Commented [A16]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

3.5 Analisis Performa Dataset IndoNLU

Pada dataset IndoNLU, metode Cosine Similarity dengan pendekatan Distributed Bag of Words (DBOW) atau DM = 0, epoch sebanyak 10, ukuran vektor 100, dan window 5 menunjukkan performa terbaik. Waktu yang diperlukan untuk pembuatan model sangat cepat, hanya 0.003 detik dengan proses Cosine yang memerlukan waktu 0.001 detik. Dari model tersebut, diperoleh akurasi sebesar 0.72, presisi 0.75, recall 0.65, dan nilai F-1 sebesar 0.65. Sementara itu, performa terbaik pada metode kedua yaitu metode Jaccard dicapai melalui penggunaan doc2vec Distributed Memory, dengan 100 epoch, ukuran vektor 100, dan window 15, menghasilkan akurasi 0.63, presisi 0.81, recall 0.52, dan nilai F-1 0.42. Waktu yang dibutuhkan untuk pembuatan model adalah 0.05 detik dan waktu proses Jaccard adalah 1.09 detik. Metode ketiga, yaitu Euclidean Distance, menghasilkan performa terbaiknya dengan parameter doc2vec Distributed Bag of Words (DBOW), maksimum epoch 400, ukuran vektor 100, dan window 5, mencatatkan akurasi 0.6, presisi 0.49, recall 0.5, dan nilai F-1 0.4. Waktu pelatihan untuk model ini adalah 0.12 detik dengan waktu proses Euclidean sebesar 1.09 detik. Dengan demikian, berdasarkan analisis ini, Cosine Similarity merupakan metode dengan performa terbaik untuk dataset IndoNLU. Ringkasan dari hasil pada dataset IndoNLU dapat dilihat pada Tabel 7.

Tabel 7. Ringkasan Hasil Dataset IndoNLU

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	0	10	100	5	0.72	0.75	0.65	0.65	0.003
Cosine	1	100	100	15	0.63	0.81	0.52	0.42	0.05
Euclidean	0	400	100	5	0.6	0.49	0.5	0.4	0.12

Commented [A17]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

3.6 Analisis Performa Dataset TaPaCo

Performa tertinggi dari penggunaan Cosine Similarity adalah dengan pendekatan Distributed Bag of Words (DBOW) atau DM = 0, di mana pengaturan epoch adalah 20, ukuran vektor 100, dan window 15, menghasilkan akurasi yang sangat tinggi yaitu 0.99, presisi 0.63, recall 0.89, dan nilai F-1 sebesar 0.69. Waktu pelatihan cukup efisien, hanya memerlukan 24.49 detik, meskipun waktu proses cosine tergolong lama, yaitu 227.91 detik. Di sisi lain, metode Jaccard menunjukkan performa terendah dibandingkan dua metode lainnya pada dataset ini, dengan penggunaan doc2vec Distributed Memory, epoch 400, ukuran vektor 10, dan window 15 menghasilkan akurasi 0.79, presisi 0.50, recall 0.73, dan nilai F-1 0.45, dengan waktu pelatihan 284.14 detik. Namun waktu proses Jaccard sangat cepat, yaitu 7.57 detik. Berbeda dari kedua dataset lainnya, metode Euclidean Distance menjadi metode dengan performa terbaik untuk dataset ini, dengan parameter terbaik meliputi doc2vec Distributed Bag of Words, epoch 400, ukuran vektor 10, dan window 5, yang menghasilkan akurasi 0.99, presisi 0.99, recall 0.77, dan nilai F-1 0.85. Waktu pelatihan yang dibutuhkan adalah 188.11 detik dengan waktu proses Euclidean sebesar 32.13 detik. Performa superior Euclidean Distance ini menandakan keefektifan pembentukan vektor doc2vec untuk dokumen dengan kalimat yang relatif singkat, menjadikannya pilihan terbaik untuk dataset ini. Oleh karena itu, pentingnya pemilihan metode dan parameter yang tepat dalam memaksimalkan performa analisis teks, terutama dalam konteks dataset dengan karakteristik tertentu. Ringkasan hasil dataset TaPaCo dilihat pada Tabel 8.

Tabel 8. Ringkasan Hasil Dataset TaPaCo

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	0	20	100	5	0.99	0.63	0.89	0.69	24.49
Cosine	1	400	10	15	0.79	0.50	0.73	0.45	284.18
Euclidean	0	400	10	5	0.99	0.99	0.77	0.85	188.11

Commented [A18]: Tabel diberikan penomoran dan judul min 2 kata. Tabel jangan di jadikan gambar. Setiap Tabel diberikan penjelasan detail dan mengkaitkan penomoran tabel pada isi penjelasan yang dilakukan. Sebelum tabel harus di berikan kalimat pengantar.

4. KESIMPULAN

Commented [A19]: Penulisan kesimpulan cukup bagus. Kesimpulan harus berisi satu paragraph, tidak menggunakan point, sudah berisi pernyataan akhir, hasil/temuan dari penelitian yang dilakukan.

Kesimpulan dari studi ini menunjukkan bahwa model berhasil membedakan antara dua dokumen yang identik atau berbeda. Performa paling unggul dengan akurasi 0.98, presisi 0.84, recall 0.95, dan skor F-1 0.89, dengan model dibentuk dalam waktu 10.56 detik menggunakan data dari Google News. Secara umum, Cosine Similarity menunjukkan hasil yang lebih unggul dibandingkan dengan Jaccard Coefficient dan Euclidean Distance berdasarkan penilaian performa rata-rata dari seluruh dataset. Konfigurasi parameter yang paling efektif untuk model dalam penelitian ini termasuk penggunaan doc2vec Distributed Memory (DM = 1) dengan 400 epoch, ukuran vektor 10, dan jendela 15 untuk Jaccard, serupa dengan Euclidean Distance, walaupun untuk Euclidean

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

jendela terbaik adalah 5. Sedangkan konfigurasi terbaik untuk Cosine Similarity adalah Distributed Bag of Words (DM = 0) dengan 10 epoch, ukuran vektor 100, dan jendela 5. Hal ini menunjukkan bagaimana berbagai parameter berpengaruh berbeda pada setiap metode. Analisis ini dilakukan menggunakan tiga dataset, dengan jumlah kata dalam Google News sebanyak 71085, di IndoNLU 6767, dan di TaPaCo 7889, dimana dataset Google News menghasilkan performa terbaik karena diversitas dan kejelasan konten serta tema dalam dokumen-dokumennya.

Rekomendasi untuk penelitian selanjutnya meliputi eksplorasi berbagai metode untuk menilai kemiripan antar dokumen dan variasi parameter yang dapat menghasilkan output yang berbeda, seperti pemanfaatan Fasttext dan BERT. Penggunaan dataset yang tidak seimbang antara kelas berita yang sama dan berbeda, yang berkontribusi pada presisi rendah. Oleh karena itu, diusulkan untuk mengembangkan dataset yang lebih seimbang, kaya, dan beragam. Disarankan juga untuk memanfaatkan dataset yang lebih baik dan lebih baru untuk mencerminkan kondisi saat ini dengan lebih akurat. Untuk dokumen dengan volume data yang lebih rendah, pertimbangan penggantian doc2vec dengan algoritma lain bisa dilakukan, begitu juga dengan proses pengukuran similaritas yang dapat diadaptasi dengan metode lain.

Commented [A20]: Sebaiknya buat dalam satu paragraf.

REFERENCES

Commented [A21]: Penulisan referensi cukup bagus. Isi Referensi sudah menggunakan Soft Referensi Ilmiah, dengan Format IEEE, Jumlah referensi yang dijadikan acuan pustaka sudah memenuhi batas minimal, untuk pustaka primer sebanyak 80% sumber referensi dari penelitian terkait sudah terpenuhi dan termutakhir 5-8 tahun terakhir.

[1] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.09129

[2] T. A. W. Tyas, Z. K. A. Baizal, and R. Dharayani, "Tourist Places Recommender System Using Cosine Similarity and Singular Value Decomposition Methods," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 4, p. 1201, Oct. 2021, doi: 10.30865/mib.v5i4.3151.

[3] I. Mawanta, T. S. Gunawan, and W. Wanayumini, "Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 726, Apr. 2021, doi: 10.30865/mib.v5i2.2935.

[4] R. Jasmi, Z. K. A. Baizal, and D. Richasdy, "Question Answering Chatbot using Ontology for History of the Sumedang Larang Kingdom using Cosine Similarity as Similarity Measure," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2040, Oct. 2022, doi: 10.30865/mib.v6i4.4530.

[5] L. Mayola, M. Hafizh, and D. Marse Putra, "Algoritma Jaccard Similarity untuk Deteksi Kemiripan Judul Disertasi dengan Pendekatan Variasi Stop Word Removal," vol. 8, no. 1, pp. 477–487, 2024, doi: 10.30865/mib.v8i1.7109.

[6] S. Pawestri, "Analisis Perbandingan Metode Jaccard Coefficient dan Cosine Similarity untuk Kemiripan Teks Bahasa Indonesia," Tesis, Universitas Gadjah Mada, Yogyakarta, 2022. Accessed: Apr. 22, 2024. [Online]. Available: https://etd.repository.ugm.ac.id/penelitian/detail/219434

[7] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 2, pp. 329–338, 2021, doi: 10.1007/s40031-020-00501-5.

[8] I. R. Hendrawan, E. Utami, and A. D. Hartanto, "Comparison of Word2vec and Doc2vec Methods for Text Classification of Product Reviews," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2022, pp. 530–534. doi: 10.1109/ICITISEE57756.2022.10057702.

[9] B. Walek and P. Müller, "An approach for recommending relevant articles in news portal based on Doc2Vec," in *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2022, pp. 26–31. doi: 10.1109/AIKE55402.2022.00010.

[10] N. V. A. Kumar and S. Mehrotra, "A Comparative Analysis of word embedding techniques and text similarity Measures," in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 2022, pp. 1581–1585. doi: 10.1109/IC3I56241.2022.10072927.

[11] A. Mandal, K. Ghosh, S. Ghosh, and S. Mandal, "Unsupervised approaches for measuring textual similarity between legal court case reports," *Artif Intell Law (Dordr)*, vol. 29, no. 3, pp. 417–451, 2021, doi: 10.1007/s10506-020-09280-2.

[12] P. K. Reshma, S. Rajagopal, and V. L. Lajish, "A Novel Document and Query Similarity Indexing using VSM for Unstructured Documents," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 676–681. doi: 10.1109/ICACCS48705.2020.9074255.

[13] K. Iwamoto, H. Uchida, Y. Li, and Y. Nakatoh, "Automatic Text-to-sound Generation by Doc2Vec," in *Human Interaction & Emerging Technologies (IHIET 2023): Artificial Intelligence & Future Applications*, AHFE International, 2023. doi: 10.54941/ahfe1004033.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

[14] K. Chen, J. Huang, Y. Cui, and W. Ren, "Research on Chinese Audio and Text Alignment Algorithm Based on AIC-FCM and Doc2Vec," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 3, Apr. 2023, doi: 10.1145/3532852.

[15] M. Alobed, A. M. M. Altrad, and Z. B. A. Bakar, "A Comparative Analysis of Euclidean, Jaccard and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring," in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2021, pp. 70–74. doi: 10.1109/CAMP51653.2021.9498119.

[16] J. Zhang, F. Wang, F. Ma, and G. Song, "Text Similarity Calculation Method Based on Optimized Cosine Distance," in *2022 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, 2022, pp. 37–39. doi: 10.1109/ICEDCS57360.2022.00015.

[17] S. Dash, T. Mohanty, S. R. Das, A. Mohanty, and R. Rautray, "PCTS: Partition Based Clustering for Text Summarization," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, 2023, pp. 1–6. doi: 10.1109/APSIT58554.2023.10201655.

[18] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," May 2014, [Online]. Available: http://arxiv.org/abs/1405.4053

[19] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Finding Similar Items," in *Mining of Massive Datasets*, 2nd ed., J. Leskovec, A. Rajaraman, and J. D. Ullman, Eds., Cambridge: Cambridge University Press, 2014, pp. 68–122. doi: DOI: 10.1017/CBO9781139924801.004.

[20] J. Han, M. Kamber, and J. Pei, "2 - Getting to Know Your Data," in *Data Mining (Third Edition)*, J. Han, M. Kamber, and J. Pei, Eds., Boston: Morgan Kaufmann, 2012, pp. 39–82. doi: https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[21] H. Parvin, H. Alizadeh, and B. Minati, "A Modification on K-Nearest Neighbor Classifier," 2010.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia

Abstrak—Kemudahan dalam mengakses informasi membawa berbagai keuntungan, termasuk kemampuan untuk mengembangkan model atau sistem yang dapat mendeteksi kemiripan antar dokumen, sistem pengecekan plagiarisme, pengelompokan teks berdasarkan tema, peringkasan otomatis, klasifikasi judul penelitian sesuai dengan topiknya, dan masih banyak lagi. Berbagai kegunaan ini membuat penelitian tentang deteksi kemiripan antar dokumen menjadi area yang penting untuk dikembangkan. Namun, studi mengenai deteksi kesamaan khusus untuk dokumen berbahasa Indonesia masih tergolong sedikit. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis komparatif terhadap kinerja Doc2Vec dibandingkan dengan Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance dalam mendeteksi kemiripan dokumen dengan teks berbahasa Indonesia. Tiga dataset digunakan dalam analisis ini, dengan dataset pertama terdiri dari 200 berita dari Google News, dataset kedua dari IndoNLU yang mempunyai 300 data, dan dataset ketiga dari TaPaCo dengan 1602 data. Temuan dari studi ini menunjukkan bahwa secara keseluruhan Cosine Similarity memiliki kinerja yang lebih baik dibandingkan Jaccard Coefficient dan Euclidean Distance dalam aspek akurasi, presisi, recall, dan nilai f-1. Di antara dataset yang diuji, Google News terbukti memiliki kinerja paling baik.

Kata Kunci: kemiripan teks Bahasa Indonesia, Doc2vec, Jaccard Coefficient, Cosine Similarity, Euclidean Distance

Abstract—*Ease of accessing information brings diverse benefits, including the ability to develop models that can detect similarities between documents, a plagiarism-checking system, automatic summarization, classification, etc. These various uses make research on similarity detection between documents a critical area to develop. However, studies about similarity detection specifically for Indonesian language documents are relatively few. Therefore, this research aims to conduct a comparative analysis of the performance of Doc2Vec compared to the Jaccard Coefficient, Cosine Similarity, and Euclidean Distance in detecting the similarity of documents with Indonesian text. Three datasets are used in this analysis, with the first dataset consisting of 200 news from Google News, the second dataset from IndoNLU, and the third dataset from TaPaCo. The findings from this study show that overall Cosine Similarity has better performance than Jaccard Coefficient and Euclidean Distance in all aspects. Among the tested datasets, Google News proved to perform the best.*

Keywords: similarity Indonesian text, Doc2vec, Jaccard Coefficient, Cosine Similarity, Euclidean Distance

1. PENDAHULUAN

Era pesatnya teknologi dan internet ini menyebabkan informasi dan data menyebar dengan sangat cepat dan mudah untuk didapatkan. Informasi dan data yang mudah didapatkan juga menyebabkan teknologi berkembang dengan pesat. Banyak manfaat dari teknologi yang pesat, salah satunya di bidang karya tulis. Hal positif dari pesatnya perkembangan teknologi adalah karya tulis dapat tersebar dengan luas dan pesat sehingga lebih populer dengan cepat. Penulis bisa mempromosikan karyanya dimana saja, kapan saja, dan tidak terbatas ruang dan waktu. Otomatisasi juga sudah mulai banyak dikembangkan untuk membantu pekerjaan manusia terutama dalam bidang karya tulis. Contohnya di bidang karya tulis adalah ada sistem yang mampu cek typo, sistem yang mampu cek grammar sebuah tulisan dalam bahasa asing, cek plagiat, cek kemiripan dua atau lebih dokumen, peringkasan paragraf otomatis, dan sebagainya. Misal saja dalam bidang pendidikan, seorang guru ingin koreksi jawaban esai dari muridnya. Namun guru tersebut juga harus cek satu-satu apakah ada esai yang sama atau saling contek atau tidak. Jika harus cek satu-satu, maka akan memakan waktu yang lama dan lebih sulit karena esai merupakan paragraf-paragraf yang panjang. Permasalahan tersebut bisa diatasi dengan cek kemiripan dokumen untuk mempermudah guru cek jawaban siswa tanpa memerlukan waktu yang lama. Namun, selain hal-hal positif, terdapat juga hal negatif yang diakibatkan dari cepatnya pergerakan data. Salah satunya adalah plagiasi karya, tulisan, atau sesuatu untuk kepentingan sendiri. Contohnya adalah seorang yang melakukan plagiasi, mahasiswa yang hanya copy-paste tanpa mencantumkan sumber, atau siswa yang mencontek karya atau tugas teman lainnya. Hal tersebut dapat diatasi dengan membuat sistem untuk cek kemiripan kata, kalimat, atau dokumen. Konsep ini mirip dengan cek plagiasi, satu dokumen diperiksa apakah dokumen tersebut mempunyai kesamaan yang sangat tinggi oleh dokumen-dokumen yang ada pada database. Jika sebuah dokumen mempunyai kesamaan kalimat atau paragraf pada dokumen-dokumen yang ada pada database, maka dapat dikatakan dokumen tersebut hanya copy-paste dari sebuah sumber. Sehingga, sistem ini bisa dikembangkan lebih lanjut menjadi cek plagiasi, tidak hanya cek kesamaan dokumen saja.

Beberapa penelitian terdahulu yang telah dilakukan tentang cek kemiripan kata adalah metode similarity untuk kemampuan maksimum interpretasi manusia pada 2019 oleh Sitikhu dkk. Metode-metode yang dilakukannya antara lain Cosine Similarity dengan TF-IDF, Cosine Similarity dengan Word2Vec Vectors, dan Soft Cosine Similarity dengan Word2Vec Vectors [1]. Penelitian cek kemiripan kata juga dilakukan oleh beberapa peneliti untuk kemiripan antara teks atau dokumen serta objek yang bukan teks juga menggunakan metode jarak dan kemiripan dua objek seperti Cosine Similarity [2], [3], [4], Jaccard Coefficient[5], gabungan dari kedua metode tersebut [6], dan Euclidean Distance [7]. Selanjutnya peneliti pada tahun 2022, Hendrawan dkk. melakukan sebuah penelitian komparasi word2vec dan doc2vec. Penelitian tersebut menyatakan bahwa Doc2vec mempunyai kelebihan dari sisi mengukur kedekatan per kalimat. Hasil penelitian menunjukkan performa Doc2Vec lebih unggul [8]. Penggunaan Doc2Vec juga diteliti oleh beberapa penelitian seperti [9],

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

Commented [A1]: 1.Analisis dari beberapa metode similirity dilakukan untuk apa? Pendeteksiankah? Atau penilaian? Atau apa?
2.Pada judul minimal 12 kata

Commented [A2]: Permasalahan masih belum jelas, jika ini adalah permasalahan yang di maksud dalam penelitian ini, maka pernyataan ini tidak dapat dikatakan sebagai permasalahan

Commented [A3]: Sebaiknya di sebutkan besarnya akurasi, presisi, recall dan nilai f-1 yang di peroleh

Commented [A4]: Metode Similarity yang akan dibandingkan tidak di jelaskan dalam bab ini.

Commented [A5]: Tidak perlu dijelaskan perkembangan teknologi saat ini, karena pembaca juga sudah tahu itu.

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

[10], [11]. Pada penelitian [12] juga menggunakan Word2Vec yang kemudian menyarankan untuk menggunakan Doc2Vec. Selain itu, Doc2Vec juga bisa mengubah teks menjadi suara seperti pada penelitian [13], [14].

Penelitian terdahulu seperti penelitian oleh Sitikhu dan Singh menyarankan cek kemiripan antar dokumen dengan menggunakan Doc2Vec agar akurasi menjadi lebih baik. Oleh karena itu, penelitian ini melakukan komparasi metode cek kemiripan dokumen Bahasa Indonesia menggunakan metode Doc2Vec. Metode similarity yang digunakan adalah metode Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Beberapa menggunakan algoritma ini sebagai metode untuk cek kemiripan dikarenakan metode tersebut pas untuk menghitung kemiripan dua objek. Penelitian terdahulu mengenai metode ini salah satunya adalah komparasi Cosine Similarity, Jaccard Coefficient, dan Euclidean Distance pada teks bahasa Arab oleh Alobed, dkk. pada tahun 2021 [15]. Selain itu, ada penelitian mengenai ringkasan teks dengan metode similarity menggunakan Cosine Similarity [16], Jaccard Coefficient, dan Euclidean Distance[17]. Tujuan dari penelitian ini adalah komparasi performa metode-metode tersebut untuk dokumen yang menggunakan Bahasa Indonesia.

Banyak manfaat dari cek kemiripan teks dan terdapat beberapa metode untuk cek kemiripan teks dokumen dengan berbagai bahasa. Banyak pula penelitian mengenai hal ini dengan berbagai bahasa. Hal ini menunjukkan kemiripan teks banyak digunakan dan dimanfaatkan di seluruh penjuru dunia. Namun penelitian untuk teks yang menggunakan bahasa Indonesia masih sedikit jumlahnya. Oleh karena itu, pengembangan lebih lanjut mengenai cek kemiripan teks berbahasa Indonesia perlu ditingkatkan. Kontribusi penelitian ini untuk menunjang pengembangan pengolahan teks bahasa Indonesia adalah dengan dilakukannya analisis komparasi berbagai metode untuk cek kemiripan dengan teks Bahasa Indonesia. Namun, tidak semua kasus dapat dilakukan pada penelitian ini, sehingga pada penelitian ini, terdapat beberapa batasan masalah. Batasan masalah penelitian ini yaitu penelitian ini menggunakan dua dataset yang dapat diakses secara bebas yaitu IndoNLU dan TaPaCo dan mengumpulkan satu dataset berita secara manual yaitu Google News. Google News berisi 200 berita Bahasa Indonesia diambil dari bulan Juni sampai Agustus 2022 secara acak dan berita tersebut telah dikelompokkan oleh Google News. Penelitian ini juga dibatasi oleh dataset yang menggunakan Bahasa Indonesia saja.

2. METODOLOGI PENELITIAN

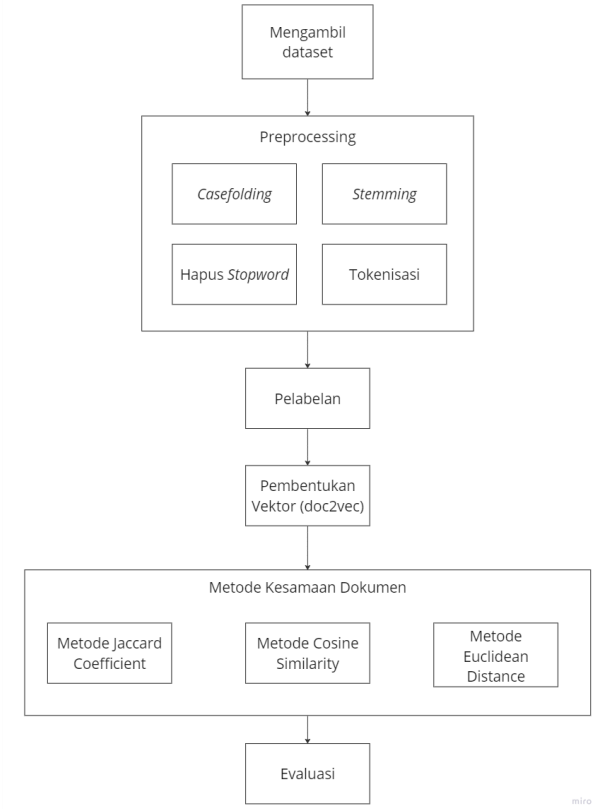
Commented [A6]: Cukup Bagus

2.1 Tahapan Penelitian

Dua buah objek yang dikatakan mirip adalah jika dua objek tersebut hampir sama atau serupa. Misal sebuah berita dari website A membicarakan mengenai kenaikan harga minyak di Indonesia, sedangkan berita dari website B juga demikian namun isi dalam berita tersebut tidak sama persis, maka kedua berita tersebut tetap dikatakan mirip karena membahas persoalan yang sama. Hal tersebut terjadi karena jika kedua berita tersebut sama persis padahal dari sumber/media yang berbeda dan tidak menyatakan sumber asli berita tersebut, bisa terkena plagiasi. Penelitian ini merupakan komparasi kemiripan atau similarity untuk teks atau dokumen berbahasa Indonesia. Metode yang digunakan untuk cek similaritas adalah Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Sedangkan untuk metode pembuatan vektor dokumen yang digunakan adalah Doc2Vec. Langkah awal dalam penelitian ini adalah mengumpulkan dataset. Kemudian dataset dimasukkan ke tahap proses preprocessing. Tahap preprocessing ada empat proses, yaitu case folding, stemming, penghapusan stopword, dan pembuatan token (Tokenisasi). Setelah preprocessing data, dataset tersebut dibentuk vektor dokumen dengan metode yang telah disebutkan. Langkah akhir adalah perhitungan kemiripan antara dua dokumen yang diikuti dengan evaluasi. Tahap evaluasi dilakukan dengan menganalisis hasil. Skema rancangan umum penelitian ini ditunjukkan pada Gambar 1.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293



Gambar 1. Tahapan Penelitian Secara Umum

2.2 *Preprocessing*

Preprocessing data yang dilakukan pada penelitian ini ada empat, yaitu case folding, stemming, menghapus stopwords, dan pembuatan token. Case folding membuat semua huruf menjadi huruf kecil. Setelah itu, proses menghapus kata imbuhan atau kata yang tidak perlu dan diikuti dengan penghapusan kata depan. Selanjutnya melakukan tokenisasi. Contoh hasil preprocessing dilihat pada Tabel 1 sebagai berikut.

Proses	Hasil *Preprocessing*
Awal	Krjogja.com - YOGYA - Badan Meteorologi Klimatologi dan Geofisika (BMKG) DIY kembali mengeluarkan peringatan dini terhadap gelombang laut di perairan Yogyakarta, Senin (04/03/2024).
Casefolding	krjogja.com - yogya - badan meteorologi klimatologi dan geofisika (bmkg) diy kembali mengeluarkan peringatan dini terhadap gelombang laut di perairan yogyakarta, senin (04/03/2024).
Stemming	krjogjacom yogya badan meteorologi klimatologi dan geofisika bmkg diy kembali keluar peringatan dini hadap gelombang laut di air yogyakarta senin 04032024
Hapus *Stopwords*	krjogjacom yogya badan meteorologi klimatologi geofisika bmkg diy kembali keluar peringatan dini hadap gelombang laut air yogyakarta senin 04032024
Tokenisasi	“krjogjacom”,“yogya”,“badan”,“meteorologi”,“klimatologi”,“geofisika”,“bmkg”,“diy ”,“kembali”,“keluar”,“peringatan”,“dini”,“hadap”,“gelombang”,“laut”,“air”, “yogyakarta”,“senin”,“04032024”

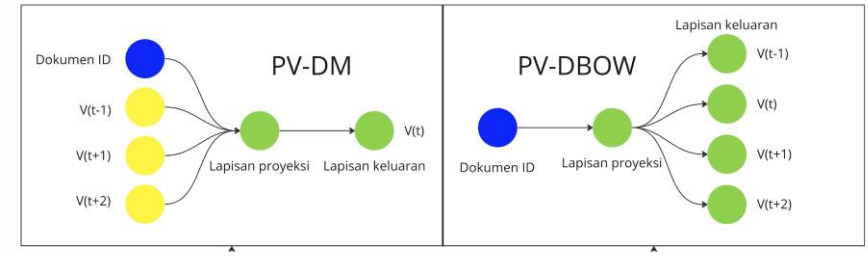
Tabel 1. Contoh *Preprocessing*

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

2.3 Doc2Vec

Doc2vec dikembangkan oleh pengembang Word2Vec juga, yaitu Le dan Mikolov pada tahun 2014 yang digunakan untuk membuat representasi numerical dari dokumen (vektor) dengan mempertimbangkan konteks atau kata-kata yang muncul [18]. Doc2vec termasuk algoritma unsupervised. Doc2vec dan word2vec mempunyai satu perbedaan, yaitu word2vec menggunakan kata (word) sedangkan doc2vec menggunakan dokumen atau paragraf untuk direpresentasikan ke dalam vektor. Doc2Vec menggunakan dua algoritma utama yaitu Paragraph Vector - Distributed Memory (PV-DM) dan Paragraph Vector - Distributed Bag of Words (PV-DBOW). Cara kerja distributed memory adalah dengan menggunakan dua input, yaitu ID dokumen dan kata-kata yang ada pada dokumen. Setelah itu akan masuk pada lapisan proyeksi yang bertugas membuat vektor kata dan vektor dokumen. Vektor-vektor tersebut masuk ke dalam fungsi yang meminimalkan perbedaan antara kada prediksi dan kata target. Setelah itu, output akan terlihat pada lapisan keluaran. Kebalikan dari DM, PV-DBOW berfokus pada kat-kata yang distribusikan dan bukan maknanya. Sehingga algoritma ini pas untuk melihat pola teks, bukan isi sebuah teks. PV-DBOW hanya mempunyai satu input, yaitu ID dokumen. ID tersebut yang nantinya menjadi vektor dokumen. Vektor yang dihasilkan dari doc2vec ini bisa digunakan untuk menghitung kemiripan antar dokumen. Berikut arsitektur Doc2Vec ditampilkan pada Gambar 2.



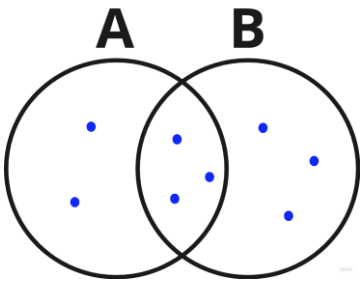
Gambar 2. Arsitektur Doc2Vec

2.4 Jaccard Coefficient

Banyak metode yang bisa digunakan untuk cek kemiripan kata atau kalimat, salah satunya dengan Jaccard Coefficient atau Jaccard Index. Jaccard Coefficient membutuhkan dua dokumen untuk dicari kemiripannya [19]. Jaccard Coefficient dari dua himpunan A dan B dapat dilihat dari persamaan nomor 1.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Sehingga dapat disimpulkan Jaccard Coefficient dari himpunan A dan B adalah jumlah elemen (kardinal) dari irisan himpunan A dan himpunan B kemudian dibagi dengan kardinal dari gabungan himpunan A dan himpunan B. Sebagai contoh dapat dilihat pada Gambar 3.



Gambar 3. Ilustrasi Jaccard Coefficient Himpunan A dan B

Gambar 3 menunjukkan kardinal dari himpunan A adalah 5, sedangkan kardinal himpunan B adalah 6. Kardinal dari irisan himpunan A dan B adalah 3. Kardinal dari gabungan himpunan A dan B adalah 8 item. Sehingga dapat disimpulkan untuk Jaccard Coefficient dari himpunan A dan B adalah 3/8. Contoh perhitungan Jaccard dapat dilihat sebagai berikut.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Terdapat 2 vektor (vektor A dan vektor B).

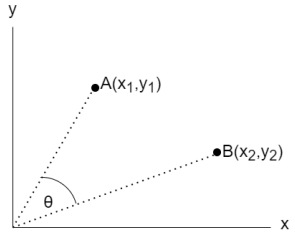
$$A = [1, 3, 6, 9], B = [1, 2, 3, 4]$$

Sehingga dengan persamaan nomor 1, Jaccard Coefficient dapat dari vektor A dan B adalah

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{2}{6} = 0.33$$

2.5 Cosine Similarity

Cosine Similarity adalah salah satu metode menghitung kemiripan antara dua kata/kalimat. Sebelum dihitung kemiripannya, kata/kalimat tersebut diubah dalam bentuk vektor. Setelah itu, Cosine Similarity akan menghitung kemiripan dua vektor dengan menghitung nilai cosinus dari sudut terkecil dari vektor tersebut. Cosine Similarity terletak antara -1 dan 1. Semakin mendekati 1 maka semakin mirip vektor tersebut [20]. Konsep cosine similarity dapat dilihat pada Gambar 4.



Gambar 4. Ilustrasi Cosine Similarity Titik A dan B

Diketahui dua vektor A dan B dengan sudut theta. Similaritas pada kedua vektor tersebut adalah dengan menggunakan persamaan nomor 2.

$$sim(A,B) = cos\ \theta = \frac{A \cdot B}{|A||B|} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{({x_1}^2 + {y_1}^2)({x_2}^2 + {y_2}^2)}} \quad (2)$$

Dimana A·B adalah dot product dari vektor A dan B. Dot product bisa dihitung dengan cara mengalikan vektor A dengan vektor B. |A| adalah panjang dari vektor A dan |B| adalah panjang dari vektor B. Contoh untuk perhitungan cosine adalah sebagai berikut. Terdapat 2 vektor (vektor A dan vektor B).

$$A = [2, 3, 4, 1], B = [1, 2, 3, 4]$$

Maka Cosine Similarity dari vektor A dan B adalah

$$Cos(A,B) = \frac{\underline{A} \cdot \underline{B}}{|\underline{A}| \cdot |\underline{B}|}$$

$$= \frac{(2 \times 1) + (3 \times 2) + (4 \times 3) + (1 \times 4)}{\sqrt{2^2 + 3^2 + 4^2 + 1^2} \cdot \sqrt{1^2 + 2^2 + 3^2 + 4^2}}$$

$$= \frac{2 + 6 + 12 + 4}{\sqrt{4 + 9 + 16 + 1} \cdot \sqrt{1 + 4 + 9 + 16}}$$

$$= \frac{24}{\sqrt{30} \cdot \sqrt{30}} = \frac{24}{30} = 0.8$$

2.6 Euclidean Distance

Euclidean Distance adalah metode untuk menghitung jarak. Semakin kecil nilai jarak, semakin mirip antara satu vektor dengan vektor lainnya. Rumus yang digunakan untuk menghitung euclidean distance dapat dilihat pada persamaan nomor 3 [21].

$$d = \sqrt{\sum_{i=1}^{n} (pi - qi)^2} \quad (3)$$

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Jarak dari vektor p ke vektor q adalah d dan n adalah jumlah elemen dari vektor p dna q. Sedangkan pi dan qi masing-masing adalah elemen ke-1 dari vektor p dan q.

3. HASIL DAN PEMBAHASAN

Commented [A7]: Cukup Bagus

Implementasi menggunakan python 3 dengan berbagai library. Pada proses implementasi, penelitian ini menggunakan parameter yang telah ditentukan. Parameter tersebut dapat dilihat pada tabel 2 sebagai berikut.

Tabel 2. Tabel Parameter

Parameter	Nilai
DM	1, 0
Max Epoch	10, 20, 100, 400
Vector Size	10, 20, 100
Window	5, 15
Min Count	1

DM adalah jenis doc2vec yang digunakan, 1 adalah Distributed Memory dan 0 adalah Distributed Bag of Words. Sedangkan max epoch adalah maksimum perulangan pada model agar lebih akurat. Kemudian vec size adalah ukuran vektor. Window adalah jarak antara kata saat ini dengan kata yang menjadi input, dan min count adalah minimal kata yang muncul. Implementasi diawali dengan mendapatkan dataset. Penelitian ini menggunakan tiga dataset. Dataset pertama adalah kumpulan berita dengan total 200 berita yang isinya diambil secara manual dari Google News pada tautan pada tautan https://news.google.com/home?hl=id&gl=ID&ceid=ID:id. Berita yang digunakan ada empat topik yang berbeda. Dataset kedua adalah dataset dari IndoNLU yang berisikan kumpulan data bisa berupa kalimat atau paragraf berbahasa Indonesia. IndoNLU bisa diakses di https://github.com/IndoNLP/indonlu. Dataset ketiga adalah dataset dari TaPaCo. TaPaCo adalah dataset yang terdiri dari kalimat dan parafrasenya dan yang diambil dari dataset TaPaCo adalah kalimat yang menggunakan bahasa Indonesia saja. Dataset TaPaCo bisa diakses di https://huggingface.co/datasets/tapaco.

Setelah mendapatkan dataset, preprocessing dilakukan untuk membersihkan data dari kata-kata yang tidak perlu agar proses menjadi lebih cepat. Setelah itu, pembentukan model. Model yang terbentuk menjadi tiga, yaitu Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Setelah model didapatkan, selanjutnya proses evaluasi. Hasil diperoleh dari melakukan tahapan evaluasi dengan membandingkan hasil label asli dengan hasil label yang diperoleh dari model. Hasil berupa akurasi, recall, precision, f1-score, dan waktu training yang didapatkan. Perhitungan evaluasi tersebut menggunakan Confusion Matrix.

3.1 Performa Model

Akurasi menandakan nilai dokumen benar diklasifikasikan dengan keseluruhan dokumen yang menjawab persoalan perbandingan dokumen mirip dan tidak mirip diprediksi benar dengan keseluruhan dokumen. Precision dan recall yang digunakan adalah rata-rata setiap kelas. Kelas pada penelitian ini ada dua, kelas 1 yang berarti kedua dokumen tersebut memiliki nilai kemiripan yang tinggi, sedangkan kelas 0 yang berarti kedua dokumen tersebut memiliki kemiripan yang rendah. Precision kelas 1 dan kelas 0 dihitung prediksi benar untuk kelas tersebut dibandingkan dengan keseluruhan data yang masuk ke dalam kelas tersebut. Setelah itu, dicari rata-rata precision untuk kelas 1 dan kelas 0. Precision untuk masing-masing kelas menjawab persoalan perbandingan berita yang terklasifikasi dengan benar dengan keseluruhan dokumen yang terklasifikasi tersebut. Kemudian recall adalah perbandingan yang terklasifikasi benar dibandingkan dengan keseluruhan yang sesungguhnya. Sedangkan, f-1 score adalah perbandingan rata-rata precision dan recall.

Hasil keseluruhan performa, didapatkan performa terbaik menggunakan Cosine Similarity karena memiliki rata-rata performa yang lebih baik daripada Jaccard Coefficient dan Euclidean Distance. Hal ini dikarenakan perhitungan Jaccard yang sederhana serta perhitungan Euclidean Distance yang kurang pas untuk kasus ini. Jaccard Coefficient membandingkan angka-angka pada vektor-vektor kemudian menghitung berapa angka yang sama dan yang berapa angka yang tidak sama. Contoh kasus dua dokumen sebenarnya mirip, namun tidak sama persis. Hal ini menyebabkan vektor juga tidak sama persis dan dapat menyebabkan kurangnya nilai kesamaan antara dua dokumen tersebut. Jika ingin performanya lebih baik, maka model harus lebih baik agar vektor yang dihasilkan memiliki vektor yang mempunyai anggota yang nilainya sama semua. Sedangkan jika menggunakan Euclidean Distance, nilai kemiripan yang dihasilkan berkisar antara 0 sampai tak hingga. Hal tersebut membuat perhitungan harus dijadikan sedemikian rupa atau dinormalisasi sehingga 0 sampai 1 saja, dengan 0 berarti tidak ada kemiripan antara dokumen-dokumen tersebut. Namun, batas-batas kemiripan lebih jelas menggunakan Cosine Similarity dan Jaccard Coefficient.

Kelebihan Cosine Similarity dibandingkan metode yang lain adalah Cosine Similarity menghitung sudut dari vektor-vektor tersebut. Kedua vektor yang dibandingkan lalu dilihat sudutnya dan dihitung nilai cosine. Jika kedua vektor tersebut didekatkan dan mempunyai sudut yang mendekati 0º, maka nilai cosine akan bernilai 1

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

yang menandakan bahwa kedua vektor tersebut sangat dekat dan dapat dikatakan kedua vektor atau dokumen tersebut mirip. Begitu juga sebaliknya, jika kedua vektor tersebut mendekati 90º, maka nilai cosine adalah 0. Sehingga dapat dikatakan kedua vektor tersebut tidak mirip. Vektor-vektor dalam kasus ini bisa lebih akurat dalam klasifikasinya dengan menggunakan Cosine Similarity karena tidak harus ada unsur yang sama, namun bisa dikatakan dokumen atau vektor yang merepresentasikan dokumen tersebut berdekatan, maka akan dinilai kedua dokumen tersebut mirip. Sebagai contoh untuk memperjelas perbedaan cara perhitungan Cosine Similarity, Jaccard Coefficient, dan Euclidean Distance, pada salah satu hasil percobaan pada dataset IndoNLU, pada sentence A dan sentence B terdapat kalimat berikut.

"Salah satu tekniknya adalah periplus , deskripsi pada pelabuhan dan daratan sepanjang garis pantai yang bisa dilihat pelaut di lepas pantai ; contoh pertamanya adalah Hanno sang Navigator dari Carthaginia dan satu lagi dari Laut Erythraea , keduanya selamat di laut menggunakan teknik periplus dengan mengenali garis pantai laut Merah dan Teluk Persi"

"Bangsa Romawi memberi sumbangan pada pemetaan karena mereka banyak menjelajahi negeri dan menambahkan teknik baru"

Kedua dokumen tersebut berlabel 0, yang artinya tidak mirip. Kedua dokumen tersebut diproses ke preprocessing dan doc2vec. Setelah perhitungan doc2vec untuk mendapatkan nilai vektor dari kedua dokumen tersebut, didapat dua vektor [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0] dan [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]. Setelah itu, vektor tersebut dihitung dengan Cosine Similarity. Perhitungan Cosine Similarity pada kasus ini adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Cosine(A,B) = \frac{(2\times 0) + \big((-6)\times 1\big) + (1\times 0) + \cdots + (6\times 0) + ((-1)\times 0) + ((-2)\times(-1))}{\sqrt{2^2 + (-6)^2 + 1^2 + \cdots + 6^2 + (-1)^2 + (-2)^2} \cdot \sqrt{0^2 + 1^2 + 0^2 + \cdots + 0^2 + 0^2 + (-1)^2}}$$

$$= \frac{0 + (-6) + 0 + 8 + 0 + 3 + 0 + 0 + 0 + 2}{\sqrt{4 + 36 + 1 + 16 + 1 + 9 + 0 + 36 + 1 + 4} \cdot \sqrt{0 + 1 + 0 + 4 + 0 + 1 + 1 + 0 + 0 + 1}}$$

$$= \frac{7}{\sqrt{108} \cdot \sqrt{8}} \approx \frac{7}{10.39 \cdot 2.83} \approx 0.24$$

Sehingga kedua dokumen ini memiliki nilai kemiripan 0.24 yang berarti tidak mirip. Mirip tidaknya menggunakan pembulatan biasa, yaitu 0.5 dikatakan kedua dokumen tersebut mirip. Namun jika 0 nilai kemiripan <0.5 , maka kedua dokumen tersebut tidak mirip. Sedangkan untuk perhitungan Jaccard Coefficient adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Jaccard(A,B) = \frac{n(\{-1.0, 0.0, 1.0, 2.0\})}{n(\{-6.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0, 4.0, 6.0\})} = \frac{4}{9} \approx 0.44$$

Jaccard Coefficient menghasilkan nilai 0.44 yang berarti tidak mirip juga walau nilai kemiripan berbeda dan Jaccard hampir mendekati nilai 0.5. Jaccard memiliki keuntungan berupa perhitungannya yang cepat karena tidak rumit. Hanya membandingkan anggota vektor mana saja yang sama dengan keseluruhan anggota vektor. Namun Jaccard juga memiliki kekurangan, perhitungannya yang sangat simple tersebut menyebabkan hitungan tidak akurat, Hanya karena berbeda beberapa anggota vektor, bukan berarti vektor tersebut tidak dekat, begitu juga sebaliknya. Sedangkan perhitungan dengan Euclidean Distance adalah sebagai berikut.

$$Vektor\ A = [2.0, -6.0, 1.0, 4.0, 1.0, 3.0, 0.0, 6.0, -1.0, -2.0]$$

$$Vektor\ B = [0.0, 1.0, -0.0, 2.0, 0.0, 1.0, 1.0, 0.0, 0.0, -1.0]$$

$$Euclidean\ (A,B) = \sqrt{(2-0)^2 + ((-6)-1)^2 + \cdots + ((-1)-0)^2 + ((-2)-(-1))^2}$$

$$Euclidean\ (A,B) = \sqrt{4 + 49 + 1 + 4 + 1 + 4 + 1 + 36 + 1 + 1} = \sqrt{102} \approx 10.1$$

Hasil menunjukkan dengan Euclidean Distance jarak yang diperoleh adalah 10.1 dan setelah melalui proses normalisasi dengan nilai dari max dari Euclidean, nilai kemiripan yang diperoleh adalah 0.64. Apabila nilai dibulatkan, nilai mendekati 1. Artinya dengan Euclidean Distance, kedua dokumen tersebut bernilai mirip. Kelebihan dari Euclidean Distance ini perhitungannya tidak lebih rumit daripada Cosine Similarity. Namun, pada kasus ini untuk mencari kemiripan dari dua vektor tidak cukup hanya mengandalkan jarak dari vektor-vektor tersebut. Pola dari vektor serta kemiripan vektor itu sendiri perlu juga untuk diperhatikan. Semua

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

parameter telah dicoba dan terdapat total 48 percobaan dari parameter yang dicoba satu-satu. Rata-rata hasil performa untuk semua dataset ditampilkan pada Tabel 3.

Tabel 3. Tabel Perbandingan Rata-rata Performa Tiga Metode

Performa	Jaccard Coefficient	Cosine Similarity	Euclidean Distance
Akurasi	0.39	0.69	0.57
Precision	0.56	0.58	0.59
Recall	0.55	0.73	0.48
F1-Score	0.29	0.49	0.47
Time (s)	8.52	23.82	20.54

Dari hasil tersebut, dapat dilihat jika rata-rata keseluruhan untuk performa Cosine lebih baik dibandingkan dengan Jaccard dan Euclidean, akan tetapi waktu proses perhitungan yang dilakukan Cosine lebih besar daripada algoritma lainnya. Hal ini dipengaruhi oleh perhitungan Cosine yang lebih kompleks dibandingkan dengan algoritma lainnya. Rata-rata tersebut berdasarkan 48 percobaan dari berbagai parameter, kemudian dicari percobaan mana saja yang pas untuk tiap-tiap metode.

Commented [A8]: Sebutkan nomor tabel yang dimaksud

3.2 Analisis Dataset

Dataset Google News yang diambil secara manual terdiri dari 200 berita dan mencakup empat tema dengan masing-masing tema diwakili oleh 40 berita. Dataset ini mempunyai jumlah total kata sebanyak 71.085 kata. Rata-rata kata dalam satu berita adalah sekitar 355 kata. Sementara itu, Dataset IndoNLU yang terdiri dari 300 dokumen atau paragraf, memiliki total kata sebanyak 6.767, dengan rata-rata jumlah kata per dokumen adalah 22 kata. Dataset TaPaCo, yang meliputi 1.602 dokumen atau paragraf, memiliki total 7.889 kata dengan rata-rata jumlah kata per dokumen berkisar antara 4 sampai 5 kata saja. Berdasarkan analisis jumlah dan keragaman kata, Dataset Google News memiliki keragaman kata dan volume yang paling tinggi dibandingkan dengan Dataset IndoNLU dan TaPaCo, yang berada pada urutan berikutnya.

Performa Dataset Google News menunjukkan hasil yang superior dibandingkan dengan dua dataset lainnya, dengan mencapai akurasi sebesar 0,98, presisi 0,84, nilai recall 0,95, dan skor F1 0,89 dalam waktu pelatihan 10,56 detik. Di sisi lain, Dataset IndoNLU menunjukkan kinerja maksimal dengan akurasi 0,72, presisi 0,75, nilai recall 0,65, dan skor F1 0,65 dalam waktu pelatihan yang sangat singkat, yaitu 0,003 detik. Dataset TaPaCo, dengan kinerja terbaiknya, menunjukkan akurasi tertinggi sebesar 0,99, presisi 0,63, nilai recall 0,89, dan skor F1 0,69 dalam waktu pelatihan 24,49 detik. Temuan ini mengindikasikan bahwa model doc2vec yang diterapkan dalam penelitian ini cenderung menghasilkan kinerja yang lebih baik pada dataset dengan keragaman kata yang lebih luas namun tetap terfokus pada satu topik, di mana satu dokumen mengandung paragraf yang terdiri dari jumlah kata yang lebih banyak.

Hal tersebut juga dipengaruhi oleh dokumen pada dataset lain mempunyai padanan kata yang sedikit. Misal pada dataset TaPaCo ada dua dokumen yang berisi kalimat “Anak perempuan mirip dengan ibunya.” dan “Buah apel tidak jatuh jauh dari pohonnya.”. Secara konteks, dua dokumen atau kalimat ini sama. Namun secara kemiripan, dua dokumen tersebut dapat dikatakan tidak mirip karena mempunyai struktur kalimat yang berbeda. Begitu juga dengan dataset IndoNLU. Berbeda dengan Google News. Google News berisi berita-berita yang tentunya jika ada sebuah berita, maka media satu dengan yang lainnya akan mirip saat menyampaikan berita tersebut secara tertulis. Jika berita yang berbeda, maka konteks dari kedua berita tersebut berbeda pula. Sehingga untuk keseluruhan performa, Google News lebih baik daripada dataset lainnya. Sedangkan waktu training, dipengaruhi oleh banyaknya dokumen dan banyaknya kata dalam suatu dokumen. Untuk informasi detail mengenai dataset dan kinerjanya dengan menggunakan model doc2vec, dapat dirujuk pada Tabel 4.

Tabel 4. Tabel Performa Terbaik Dataset

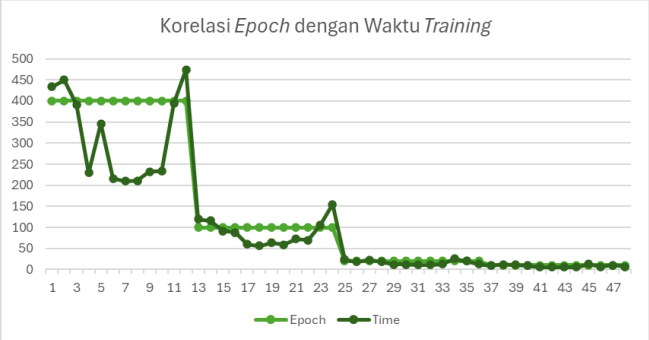
Performa	Google News	IndoNLU	TaPaCo
Banyak dokumen	200	300	1602
Total kata	71085	6767	7889
Rata-rata kata	355	22	4-5
Akurasi	0.98	0.72	0.99
Presisi	0.84	0.75	0.63
Recall	0.95	0.65	0.89
f-1 score	0.89	0.65	0.69
Waktu *training* (detik)	10.56	0.003	24.49

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

3.3 Analisis Parameter

Hasil kinerja terbaik untuk parameter setiap metode kemiripan secara umum dicapai dengan metode doc2vec 0, yang berarti menggunakan metode Distributed Bag of Words. Distributed Bag of Words terbaik diperoleh dari metode kemiripan Cosine Similarity dan Euclidean Distance. Sedangkan performa terbaik Jaccard Coefficient menggunakan doc2vec Distributed Memory. Maksimal epoch terbaik adalah 400 pada Jaccard dan Euclidean, sedangkan maksimal epoch terbaik 10 pada Cosine. Epoch menandakan banyaknya perulangan yang terjadi pada model agar memperoleh hasil yang terbaik. Cosine Similarity termasuk singkat dengan 10 epoch. Sedangkan Jaccard dan Euclidean membutuhkan 400 epoch agar hasil maksimal. Hal ini juga berpengaruh pada waktu yang dibutuhkan sebuah model untuk menjalankan tugasnya. Semakin besar epoch, semakin besar pula waktu yang dibutuhkan dalam satu kali training. Walau waktu training lebih lama daripada percobaan lainnya, namun pada Jaccard dan Euclidean menghasilkan performa yang paling baik. Korelasi epoch dengan waktu training diilustrasikan pada Gambar 5.



Gambar 5. Korelasi *Epoch* dan Waktu Training

Dengan demikian, pendekatan menggunakan Cosine dengan 10 epoch dinilai lebih efektif karena membutuhkan waktu yang singkat dengan performa terbaik. Ukuran vektor yang menghasilkan kinerja terbaik untuk semua dataset adalah 100 untuk Cosine dan 10 untuk Jaccard dan Euclidean. Ukuran vektor yang dimaksud adalah ukuran pada suatu vektor, apabila 10 maka ada 10 elemen pada satu vektor. Selain itu, ukuran window yang optimal dalam penelitian ini adalah 5 untuk Cosine dan Euclidean dan 10 untuk Jaccard. Window menunjukkan banyaknya kata yang terlibat. Parameter terbaik yang diterapkan dalam penelitian ini dilihat pada Tabel 5.

Tabel 5. Parameter Terbaik

Similarity	Doc2Vec	Epoch	Vektor	Window
Jaccard	Distributed Memory	400	10	15
Cosine	Distributed Bag of Words	10	100	5
Euclidean	Distributed Bag of Words	400	10	5

3.4 Analisis Performa Dataset Google News

Implementasi pada dataset Google News menunjukkan bahwa Cosine Similarity dengan penggunaan Distributed Memory (DM), epoch maksimal 10, ukuran vektor 100, dan window 5, menghasilkan model dengan performa terbaik untuk dataset ini. Waktu training model ini selama 10.56 detik dan waktu proses Cosine selama 1.6 detik, dengan hasil akurasi sebesar 0.98, presisi 0.84, recall 0.95, dan nilai F-1 0.89. Di sisi lain, metode Jaccard memerlukan waktu pelatihan yang lebih lama yaitu 433.61 detik dengan akurasi terbaik yang dicapai pada epoch 400, menggunakan Distributed Memory (DM), ukuran vektor 20, dan window 15, menghasilkan akurasi 0.98. Hal ini menunjukkan bahwa meskipun Jaccard mencapai akurasi yang hampir sama dengan Cosine, namun waktu yang diperlukan jauh lebih lama, dengan nilai recall dan F-1 yang tidak sebaik Cosine, sehingga kinerjanya dianggap kurang optimal. Pendekatan menggunakan metode Euclidean dengan doc2vec Distributed Memory (DM), epoch 400, ukuran vektor 10, dan window 5 mendapatkan akurasi 0.8, presisi 0.48, recall 0.42, dan nilai F-1 0.45, serta waktu proses Euclidean terlama dibanding metode yang lain yaitu 17.24 detik dan dan waktu pelatihan 391.31 detik. Berdasarkan analisis ini, Cosine Similarity terbukti sebagai metode terbaik untuk dataset ini berkat efisiensi waktu dan kinerja yang unggul. Dataset ini juga lebih pas menggunakan doc2vec Distributed Memory (DM) dibandingkan dengan Distributed Bag of Words (DBOW). Ringkasan hasil implementasi pada dataset Google News tersaji dalam Tabel 6.

Tabel 6. Ringkasan Hasil Dataset Google News

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	1	10	100	5	0.98	0.84	0.95	0.89	10.56
Cosine	1	400	20	15	0.98	0.99	0.67	0.75	433.61
Euclidean	1	400	10	5	0.80	0.48	0.42	0.45	391.31

3.5 Analisis Performa Dataset IndoNLU

Pada dataset IndoNLU, metode Cosine Similarity dengan pendekatan Distributed Bag of Words (DBOW) atau DM = 0, epoch sebanyak 10, ukuran vektor 100, dan window 5 menunjukkan performa terbaik. Waktu yang diperlukan untuk pembuatan model sangat cepat, hanya 0.003 detik dengan proses Cosine yang memerlukan waktu 0.001 detik. Dari model tersebut, diperoleh akurasi sebesar 0.72, presisi 0.75, recall 0.65, dan nilai F-1 sebesar 0.65. Sementara itu, performa terbaik pada metode kedua yaitu metode Jaccard dicapai melalui penggunaan doc2vec Distributed Memory, dengan 100 epoch, ukuran vektor 100, dan window 15, menghasilkan akurasi 0.63, presisi 0.81, recall 0.52, dan nilai F-1 0.42. Waktu yang dibutuhkan untuk pembuatan model adalah 0.05 detik dan waktu proses Jaccard adalah 1.09 detik. Metode ketiga, yaitu Euclidean Distance, menghasilkan performa terbaiknya dengan parameter doc2vec Distributed Bag of Words (DBOW), maksimum epoch 400, ukuran vektor 100, dan window 5, mencatatkan akurasi 0.6, presisi 0.49, recall 0.5, dan nilai F-1 0.4. Waktu pelatihan untuk model ini adalah 0.12 detik dengan waktu proses Euclidean sebesar 1.09 detik. Dengan demikian, berdasarkan analisis ini, Cosine Similarity merupakan metode dengan performa terbaik untuk dataset IndoNLU. Ringkasan dari hasil pada dataset IndoNLU dapat dilihat pada Tabel 7.

Tabel 7. Ringkasan Hasil Dataset IndoNLU

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	0	10	100	5	0.72	0.75	0.65	0.65	0.003
Cosine	1	100	100	15	0.63	0.81	0.52	0.42	0.05
Euclidean	0	400	100	5	0.6	0.49	0.5	0.4	0.12

3.6 Analisis Performa Dataset TaPaCo

Performa tertinggi dari penggunaan Cosine Similarity adalah dengan pendekatan Distributed Bag of Words (DBOW) atau DM = 0, di mana pengaturan epoch adalah 20, ukuran vektor 100, dan window 15, menghasilkan akurasi yang sangat tinggi yaitu 0.99, presisi 0.63, recall 0.89, dan nilai F-1 sebesar 0.69. Waktu pelatihan cukup efisien, hanya memerlukan 24.49 detik, meskipun waktu proses cosine tergolong lama, yaitu 227.91 detik. Di sisi lain, metode Jaccard menunjukkan performa terendah dibandingkan dua metode lainnya pada dataset ini, dengan penggunaan doc2vec Distributed Memory, epoch 400, ukuran vektor 10, dan window 15 menghasilkan akurasi 0.79, presisi 0.50, recall 0.73, dan nilai F-1 0.45, dengan waktu pelatihan 284.14 detik. Namun waktu proses Jaccard sangat cepat, yaitu 7.57 detik. Berbeda dari kedua dataset lainnya, metode Euclidean Distance menjadi metode dengan performa terbaik untuk dataset ini, dengan parameter terbaik meliputi doc2vec Distributed Bag of Words, epoch 400, ukuran vektor 10, dan window 5, yang menghasilkan akurasi 0.99, presisi 0.99, recall 0.77, dan nilai F-1 0.85. Waktu pelatihan yang dibutuhkan adalah 188.11 detik dengan waktu proses Euclidean sebesar 32.13 detik. Performa superior Euclidean Distance ini menandakan keefektifan pembentukan vektor doc2vec untuk dokumen dengan kalimat yang relatif singkat, menjadikannya pilihan terbaik untuk dataset ini. Oleh karena itu, pentingnya pemilihan metode dan parameter yang tepat dalam memaksimalkan performa analisis teks, terutama dalam konteks dataset dengan karakteristik tertentu. Ringkasan hasil dataset TaPaCo dilihat pada Tabel 8.

Tabel 8. Ringkasan Hasil Dataset TaPaCo

Metode Similarity	Parameter				Hasil				
	DM	max epoch	vec size	window	Akurasi	Presisi	*Recall*	*F1 score*	Waktu *Train (s)*
Jaccard	0	20	100	5	0.99	0.63	0.89	0.69	24.49
Cosine	1	400	10	15	0.79	0.50	0.73	0.45	284.18
Euclidean	0	400	10	5	0.99	0.99	0.77	0.85	188.11

4. KESIMPULAN

Commented [A9]: Kesimpulan berisi satu paragraf

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Kesimpulan dari studi ini menunjukkan bahwa model berhasil membedakan antara dua dokumen yang identik atau berbeda. Performa paling unggul dengan akurasi 0.98, presisi 0.84, recall 0.95, dan skor F-1 0.89, dengan model dibentuk dalam waktu 10.56 detik menggunakan data dari Google News. Secara umum, Cosine Similarity menunjukkan hasil yang lebih unggul dibandingkan dengan Jaccard Coefficient dan Euclidean Distance berdasarkan penilaian performa rata-rata dari seluruh dataset. Konfigurasi parameter yang paling efektif untuk model dalam penelitian ini termasuk penggunaan doc2vec Distributed Memory (DM = 1) dengan 400 epoch, ukuran vektor 10, dan jendela 15 untuk Jaccard, serupa dengan Euclidean Distance, walaupun untuk Euclidean jendela terbaik adalah 5. Sedangkan konfigurasi terbaik untuk Cosine Similarity adalah Distributed Bag of Words (DM = 0) dengan 10 epoch, ukuran vektor 100, dan jendela 5. Hal ini menunjukkan bagaimana berbagai parameter berpengaruh berbeda pada setiap metode. Analisis ini dilakukan menggunakan tiga dataset, dengan jumlah kata dalam Google News sebanyak 71085, di IndoNLU 6767, dan di TaPaCo 7889, dimana dataset Google News menghasilkan performa terbaik karena diversitas dan kejelasan konten serta tema dalam dokumen-dokumennya.

Rekomendasi untuk penelitian selanjutnya meliputi eksplorasi berbagai metode untuk menilai kemiripan antar dokumen dan variasi parameter yang dapat menghasilkan output yang berbeda, seperti pemanfaatan Fasttext dan BERT. Penggunaan dataset yang tidak seimbang antara kelas berita yang sama dan berbeda, yang berkontribusi pada presisi rendah. Oleh karena itu, diusulkan untuk mengembangkan dataset yang lebih seimbang, kaya, dan beragam. Disarankan juga untuk memanfaatkan dataset yang lebih baik dan lebih baru untuk mencerminkan kondisi saat ini dengan lebih akurat. Untuk dokumen dengan volume data yang lebih rendah, pertimbangan penggantian doc2vec dengan algoritma lain bisa dilakukan, begitu juga dengan proses pengukuran similaritas yang dapat diadaptasi dengan metode lain.

REFERENCES

Commented [A10]: Gunakan penelitian terbaru (minimal 5 tahun terakhir)

[1] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.09129

[2] T. A. W. Tyas, Z. K. A. Baizal, and R. Dharayani, "Tourist Places Recommender System Using Cosine Similarity and Singular Value Decomposition Methods," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 4, p. 1201, Oct. 2021, doi: 10.30865/mib.v5i4.3151.

[3] I. Mawanta, T. S. Gunawan, and W. Wanayumini, "Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 726, Apr. 2021, doi: 10.30865/mib.v5i2.2935.

[4] R. Jasmi, Z. K. A. Baizal, and D. Richasdy, "Question Answering Chatbot using Ontology for History of the Sumedang Larang Kingdom using Cosine Similarity as Similarity Measure," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2040, Oct. 2022, doi: 10.30865/mib.v6i4.4530.

[5] L. Mayola, M. Hafizh, and D. Marse Putra, "Algoritma Jaccard Similarity untuk Deteksi Kemiripan Judul Disertasi dengan Pendekatan Variasi Stop Word Removal," vol. 8, no. 1, pp. 477–487, 2024, doi: 10.30865/mib.v8i1.7109.

[6] S. Pawestri, "Analisis Perbandingan Metode Jaccard Coefficient dan Cosine Similarity untuk Kemiripan Teks Bahasa Indonesia," Tesis, Universitas Gadjah Mada, Yogyakarta, 2022. Accessed: Apr. 22, 2024. [Online]. Available: https://etd.repository.ugm.ac.id/penelitian/detail/219434

[7] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 2, pp. 329–338, 2021, doi: 10.1007/s40031-020-00501-5.

[8] I. R. Hendrawan, E. Utami, and A. D. Hartanto, "Comparison of Word2vec and Doc2vec Methods for Text Classification of Product Reviews," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2022, pp. 530–534. doi: 10.1109/ICITISEE57756.2022.10057702.

[9] B. Walek and P. Müller, "An approach for recommending relevant articles in news portal based on Doc2Vec," in *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2022, pp. 26–31. doi: 10.1109/AIKE55402.2022.00010.

[10] N. V. A. Kumar and S. Mehrotra, "A Comparative Analysis of word embedding techniques and text similarity Measures," in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 2022, pp. 1581–1585. doi: 10.1109/IC3I56241.2022.10072927.

[11] A. Mandal, K. Ghosh, S. Ghosh, and S. Mandal, "Unsupervised approaches for measuring textual similarity between legal court case reports," *Artif Intell Law (Dordr)*, vol. 29, no. 3, pp. 417–451, 2021, doi: 10.1007/s10506-020-09280-2.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

[12] P. K. Reshma, S. Rajagopal, and V. L. Lajish, "A Novel Document and Query Similarity Indexing using VSM for Unstructured Documents," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 676–681. doi: 10.1109/ICACCS48705.2020.9074255.

[13] K. Iwamoto, H. Uchida, Y. Li, and Y. Nakatoh, "Automatic Text-to-sound Generation by Doc2Vec," in *Human Interaction & Emerging Technologies (IHIET 2023): Artificial Intelligence & Future Applications*, AHFE International, 2023. doi: 10.54941/ahfe1004033.

[14] K. Chen, J. Huang, Y. Cui, and W. Ren, "Research on Chinese Audio and Text Alignment Algorithm Based on AIC-FCM and Doc2Vec," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 3, Apr. 2023, doi: 10.1145/3532852.

[15] M. Alobed, A. M. M. Altrad, and Z. B. A. Bakar, "A Comparative Analysis of Euclidean, Jaccard and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring," in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2021, pp. 70–74. doi: 10.1109/CAMP51653.2021.9498119.

[16] J. Zhang, F. Wang, F. Ma, and G. Song, "Text Similarity Calculation Method Based on Optimized Cosine Distance," in *2022 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, 2022, pp. 37–39. doi: 10.1109/ICEDCS57360.2022.00015.

[17] S. Dash, T. Mohanty, S. R. Das, A. Mohanty, and R. Rautray, "PCTS: Partition Based Clustering for Text Summarization," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, 2023, pp. 1–6. doi: 10.1109/APSIT58554.2023.10201655.

[18] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," May 2014, [Online]. Available: http://arxiv.org/abs/1405.4053

[19] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Finding Similar Items," in *Mining of Massive Datasets*, 2nd ed., J. Leskovec, A. Rajaraman, and J. D. Ullman, Eds., Cambridge: Cambridge University Press, 2014, pp. 68–122. doi: DOI: 10.1017/CBO9781139924801.004.

[20] J. Han, M. Kamber, and J. Pei, "2 - Getting to Know Your Data," in *Data Mining (Third Edition)*, J. Han, M. Kamber, and J. Pei, Eds., Boston: Morgan Kaufmann, 2012, pp. 39–82. doi: https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[21] H. Parvin, H. Alizadeh, and B. Minati, "A Modification on K-Nearest Neighbor Classifier," 2010.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**