Pengelompokan Dokumen Menggunakan Algoritma Doc2Vec Dan HDBSCAN Untuk Deteksi Plagiarisme

Bondan Tiur Mahendra [a,1,*], Budi Santoso [b,2], Ratna Nur Tiara Shanty [c,3]

[a,b,c] Teknik Informatika, Universitas Dr Soetomo, Indonesia
[1] bm333936@gmail.com*; [2] budi.santoso@unitomo.ac.id; [3] ratna.nurtiara03@gmail.com
* Penulis Korespondensi

ABSTRAK

Plagiarisme menjadi tantangan serius dalam lingkungan akademik karena ketersediaan konten digital yang mudah diakses. Cara deteksi plagiarisme yang biasa digunakan, yaitu dengan membandingkan kalimat secara langsung, sering kali bisa dihindari dengan cara mengubah kalimat atau melakukan perubahan kecil pada teks. Penelitian ini bertujuan membuat sistem deteksi plagiarisme yang lebih baik dengan menggunakan algoritma Doc2Vec dan HDBSCAN untuk mengelompokkan dokumen. Metode ini bekerja dengan mengubah dokumen menjadi bentuk vektor yang memiliki makna yang dalam menggunakan Doc2Vec, kemudian mengelompokkan dokumen yang memiliki konten serupa dengan HDBSCAN. Kelebihan HDBSCAN adalah mampu mengklasifikasikan dokumen asli sebagai data yang tidak relevan, sehingga meningkatkan ketepatan hasil deteksi. Uji coba dilakukan pada data esai siswa dan menunjukkan bahwa pendekatan ini mampu mengelompokkan dokumen dengan isi yang mirip, dengan skor Silhouette sebesar 0,6653 yang menunjukkan pemisahan kelompok yang baik. Penelitian ini berkontribusi dalam menyediakan alat deteksi plagiarisme yang lebih andal dan bernuansa, mampu mendeteksi kesamaan ide, bukan hanya kata.

Riwayat Artikel
Diterima 1 September 2025
Diperbaiki 4 Oktober 2025
Diterbitkan 25 Oktober 2025

Kata Kunci
Pengelompokan Dokumen
Doc2Vec
HDBSCAN
Deteksi Plagiarisme

1. Pendahuluan

Perkembangan teknologi di era digital telah menghadirkan aplikasi pembelajaran online sebagai metode populer yang memfasilitasi akses materi pembelajaran dan interaksi antara dosen dan mahasiswa tanpa batasan geografis [1], [2]. Keunggulan aplikasi pembelajaran daring membawa tantangan baru bagi dosen dalam mengawal integritas akademik dan menjaga kualitas karya ilmiah mahasiswa [3]. Minimnya pengawasan terhadap kejujuran akademik dapat mengancam keutuhan karya ilmiah dan menghambat perkembangan keilmuan di perguruan tinggi.

Menjamurnya plagiarisme merupakan masalah serius dalam etika akademik, dimana karya orang lain diambil tanpa mencantumkan sumber persisnya [4]. Pembelajaran daring melalui aplikasi digital membuka peluang lebih besar bagi mahasiswa untuk menjiplak karena kemudahan akses dan berbagi informasi dari berbagai sumber [5]. Tersedianya konten digital menyebabkan mahasiswa cenderung mengabaikan prinsip amati, tiru, dan modifikasi dalam mengembangkan karya ilmiah. Suatu studi bahkan menemukan bahwa 18% artikel memiliki kasus plagiarisme, dan jumlah tersebut cenderung lebih tinggi pada studi dengan sampel kecil atau yang menggunakan kriteria deteksi plagiarisme yang lebih ketat [6]. Kondisi ini menunjukkan urgensi untuk mengembangkan mekanisme yang efektif dan andal dalam mendeteksi kemiripan teks. Untuk mencegah dan mengurangi plagiarisme yang dapat merugikan proses pembelajaran, penelitian ini bertujuan untuk membangun sistem pendeteksi plagiarisme berbasis machine learning yang mampu mengidentifikasi kemiripan dokumen dengan mengelompokkan dokumen yang diunggah.

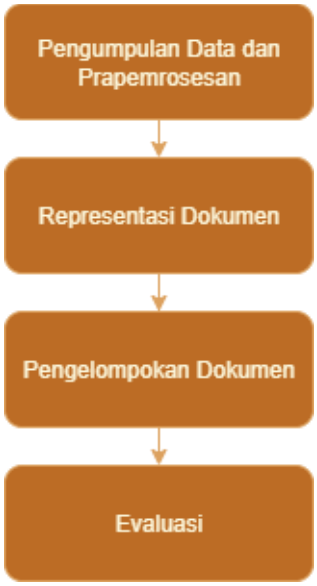
Meskipun telah banyak penelitian sebelumnya yang berupaya mengatasi tantangan ini, sebagian besar masih mengandalkan pendekatan berbasis kemiripan string seperti algoritma Jaccard Similarity atau n-gram [7]. Metode-metode ini efektif untuk mendeteksi plagiarisme langsung (*word-for-word*)

namun rentan terhadap manipulasi seperti parafrase, penyisipan sinonim, atau perubahan struktur kalimat, yang sering disebut plagiarisme semantik [8]. Beberapa studi lain telah mulai menggunakan metode embedding seperti Word2Vec, tetapi seringkali terbatas pada analisis kemiripan kata per kata, bukan makna keseluruhan dokumen [9]. Oleh karena itu, penelitian terdahulu belum secara optimal menangkap kemiripan ide atau konsep. Berbeda dengan penelitian tersebut, studi ini menawarkan pendekatan yang lebih maju dengan mengintegrasikan Doc2Vec yang dapat menangkap representasi semantik keseluruhan dokumen, serta HDBSCAN yang mampu mengelompokkan data kompleks secara adaptif. Kombinasi ini bertujuan untuk mengatasi keterbatasan metode sebelumnya dan menghasilkan deteksi plagiarisme yang lebih akurat dan menyeluruh.

Pendekatan untuk memecahkan tantangan ini membutuhkan studi mendalam tentang Pemrosesan Bahasa Alami (NLP) dan teknik pembelajaran mesin. Doc2Vec, perpanjangan dari algoritma Word2Vec, adalah metode penting untuk merepresentasikan dokumen dengan mengubahnya menjadi vektor numerik yang padat sambil mempertahankan konteks semantik teks [10]. Vektor-vektor ini kemudian dapat diproses dengan algoritma pengelompokan seperti *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Algoritma HDBSCAN dipilih karena kemampuannya untuk mengidentifikasi kelompok dengan bentuk dan kepadatan tidak beraturan, dan untuk secara otomatis mengabaikan pencilan [11], yang sangat cocok untuk data dokumen yang kompleks. Kombinasi ini memungkinkan pengelompokan dokumen serupa secara akurat, yang mengarah pada deteksi plagiarisme yang lebih efisien dan tepat. Penelitian ini diharapkan dapat menghasilkan sistem pendeteksian plagiarisme yang inovatif dan efektif, khususnya untuk konteks pembelajaran daring. Manfaat utamanya antara lain meningkatkan integritas akademik dengan menyediakan alat yang andal bagi dosen, meningkatkan efisiensi penilaian dengan mengurangi beban kerja manual, meningkatkan kualitas pendidikan dengan mendorong mahasiswa menghasilkan karya orisinal, dan memberikan kontribusi ilmiah dengan menawarkan wawasan baru tentang efektivitas penggabungan Doc2Vec dan HDBSCAN untuk deteksi plagiarisme.

2. Metode

Penelitian ini menggunakan kerangka kerja pengelompokan dokumen akademik untuk mendukung pendeteksian plagiarisme. Kerangka kerja ini mengintegrasikan Doc2Vec untuk representasi dokumen, HDBSCAN untuk pengelompokan tanpa pengawasan, dan dua pendekatan evaluasi: Skor Silhouette sebagai ukuran kuantitatif kualitas *cluster* dan visualisasi UMAP sebagai penilaian kualitatif keterpisahan kelompok dokumen.



Gambar 1 Diagram Alur Penelitian

Diagram alur menunjukkan tahapan penelitian yang berurutan. Prosesnya dimulai dengan pendataan berupa tugas siswa, dilanjutkan dengan langkah-langkah preprocessing seperti pembersihan, *case folding*, tokenisasi, penghapusan *stopword*, dan *stemming*. Setelah itu, dokumen diubah menjadi representasi vektor padat menggunakan Doc2Vec, yang memungkinkan kemiripan semantik ditangkap secara numerik. Penyematan yang dihasilkan kemudian dikelompokkan dengan HDBSCAN untuk mengidentifikasi grup dokumen dan pencilan. Terakhir, hasil pengelompokan dievaluasi baik secara kuantitatif, menggunakan skor Silhouette, maupun kualitatif, melalui visualisasi UMAP. Kerangka kerja diakhiri dengan interpretasi hasil, dimana *cluster* yang teridentifikasi dianalisis untuk kasus plagiarisme potensial dan deteksi orisinalitas.

2.1. Pengumpulan Data dan Prapemrosesan

Kumpulan data tersebut terdiri dari dokumen akademik berupa tugas siswa. Untuk memperkaya keragaman, beberapa dokumen diparafrasekan secara manual dan menggunakan alat parafrase, mengikuti praktik yang ada dalam tolok ukur deteksi plagiarisme. Tahap prapemrosesan melibatkan pembersihan, *case folding*, tokenisasi, penghapusan *stopword*, dan *stemming* [12]. Langkah-langkah ini memastikan konsistensi dalam korpus teks dan mengurangi noise sebelum menyematkan ke dalam ruang vektor.

2.2. Representasi Dokumen

Doc2Vec adalah metode yang mengajarkan komputer untuk mengubah dokumen teks menjadi vektor berbasis angka, yang dapat digunakan dalam berbagai aktivitas pembelajaran mesin seperti menyortir teks, mengelompokkan item serupa, dan menyarankan konten yang relevan [10]. Doc2Vec adalah peningkatan dari metode Word2Vec yang merepresentasikan seluruh dokumen sebagai vektor, bukan hanya satu kata, dan pendekatan ini telah ditemukan untuk menciptakan fitur yang lebih baik untuk mengkategorikan produk secara otomatis [10], [13]. Untuk mendapatkan arti kata, model Doc2Vec mencoba membuatnya lebih mungkin menebak kata berdasarkan kata-kata di sekitarnya dan makna keseluruhan dokumen. Ada dua cara untuk melatih model Doc2Vec: Memori Terdistribusi (DM) dan Kumpulan Kata Terdistribusi (DBOW). Dalam penelitian ini, model DM dipilih karena mempertahankan urutan kata dan konteks semantik secara lebih efektif. Model dilatih dengan parameter berikut:

1. Vector size = 100 dimensions,
2. Window size = 5 words,
3. Minimum count = 2,
4. Epochs = 10 iterations,
5. Workers = 4.

Ini secara formal diungkapkan dengan memaksimalkan probabilitas log dengan fungsi pada persamaan 1 [14]:

$$L \sum_{\omega \in D} \sum_{c \in C(\omega)} logp(c \vee \omega, D) \quad (1)$$

Dimana L adalah fungsi kemungkinan yang akan dimaksimalkan, ω adalah kata target, D adalah vektor dokumen, dan C(ω) representasi konteks kata dari ω. Penyematan yang dihasilkan menyandikan kesamaan semantik sehingga dokumen dengan makna yang tumpang tindih diposisikan lebih dekat dalam ruang vektor.

2.3. Pengelompokan Dokumen

Penyematan dokumen dikelompokkan bersama menggunakan HDBSCAN, sejenis metode pengelompokan yang melihat seberapa dekat titik data satu sama lain. Ini menciptakan struktur *cluster* seperti pohon dan kemudian memotongnya untuk menemukan pengelompokan yang paling andal, memungkinkan deteksi *cluster* dengan kepadatan yang bervariasi dan secara otomatis

menangani pencilan (*noise*) [15]. HDBSCAN berbeda dari K-Means karena tidak perlu memutuskan berapa banyak grup yang akan ada sebelum memulai, dan HDBSCAN dapat menangani *cluster* yang tidak semuanya berukuran sama atau tersebar [16]. Ukuran kunci dalam HDBSCAN adalah jarak jangkauan bersama [15] ditunjukkan pada persamaan 2.

$$d_{mreach}(a,b) = max\big(core_k(a), core_k(b), d(a,b)\big) \quad (2)$$

Di mana $core_k(x)$ adalah jarak dari titik *x* ke tetangga terdekat ke-k dan d(a, b) adalah jarak Euclidean antara a dan b. Berdasarkan metrik ini, HDBSCAN membangun pohon rentang jarak minimum, mengidentifikasi kelompok dengan kepadatan yang bervariasi, dan memberi label titik dengan kepadatan rendah sebagai derau. Proses pengelompokan mengikuti langkah-langkah berikut:

1. Buat penyematan Doc2Vec untuk setiap dokumen.
2. Hitung jarak berpasangan antara vektor dokumen.
3. Terapkan HDBSCAN untuk membentuk *cluster* padat dan menandai anomali sebagai noise.

2.4. Evaluasi

Kinerja pengelompokan dievaluasi dengan menggunakan pendekatan kuantitatif dan kualitatif:

2.4.1.Skor Silhouette (Kuantitatif)

Skor Silhouette adalah standar yang digunakan untuk menilai efektivitas metode pengelompokan dengan menentukan seberapa tepat setiap titik data termasuk dalam klaster yang ditentukan dalam kaitannya dengan klaster lain [17]. Skor ini ditemukan dengan memeriksa seberapa mirip setiap titik data dengan grupnya dan seberapa berbedanya dengan grup lain [17]. Untuk setiap dokumen i, nilai Silhouette didefinisikan pada persamaan 3 [18].

$$s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}} \quad (3)$$

Dimana a(*i*) adalah jarak intra-*cluster* rata-rata *i* dan b(*i*) adalah jarak rata-rata minimum dari *i* ke titik-titik di *cluster* lain. Skor Silhouette keseluruhan diperoleh dengan rata-rata semua nilai s(*i*) di seluruh dokumen. Skor mendekati 1 berarti *cluster* dipisahkan dengan jelas dan masuk akal, skor mendekati 0 berarti *cluster* tercampur, dan skor negatif menunjukkan bahwa *cluster* ditetapkan secara salah. Pengukuran ini membantu memeriksa apakah HDBSCAN dapat mengelompokkan dokumen serupa dengan benar dan menghilangkan kebisingan.

2.4.2.UMAP (Kualitatif)

UMAP adalah metode yang kuat untuk mengurangi jumlah dimensi, dan telah menjadi sangat populer dalam pembelajaran mesin dan visualisasi data [19]. Untuk memeriksa seberapa baik *cluster* dipisahkan, UMAP digunakan untuk menampilkan penyematan Doc2Vec dalam dua dimensi. Dalam proyeksi UMAP, kluster yang tampak kompak dan terpisah dengan jelas menunjukkan pengelompokan yang efektif dari dokumen yang serupa secara semantik, sementara titik-titik yang tersebar jauh dari kluster biasanya mewakili pencilan atau kiriman unik. Visualisasi ini melengkapi Skor Siluet dengan menawarkan cara intuitif untuk mengonfirmasi apakah grup yang diidentifikasi selaras dengan struktur semantik kumpulan data.

3. Hasil dan Pembahasan

Berdasarkan analisis pengelompokan menggunakan algoritma Doc2Vec dan HDBSCAN, dokumen yang diserahkan dibagi menjadi tiga kategori utama: Cluster 0, Cluster 1, Cluster 2 dan Noise. Pengelompokan ini berfungsi untuk mengidentifikasi kesamaan tersembunyi atau pola semantik di antara dokumen. Tabel 1 memberikan ringkasan distribusi pengelompokan.

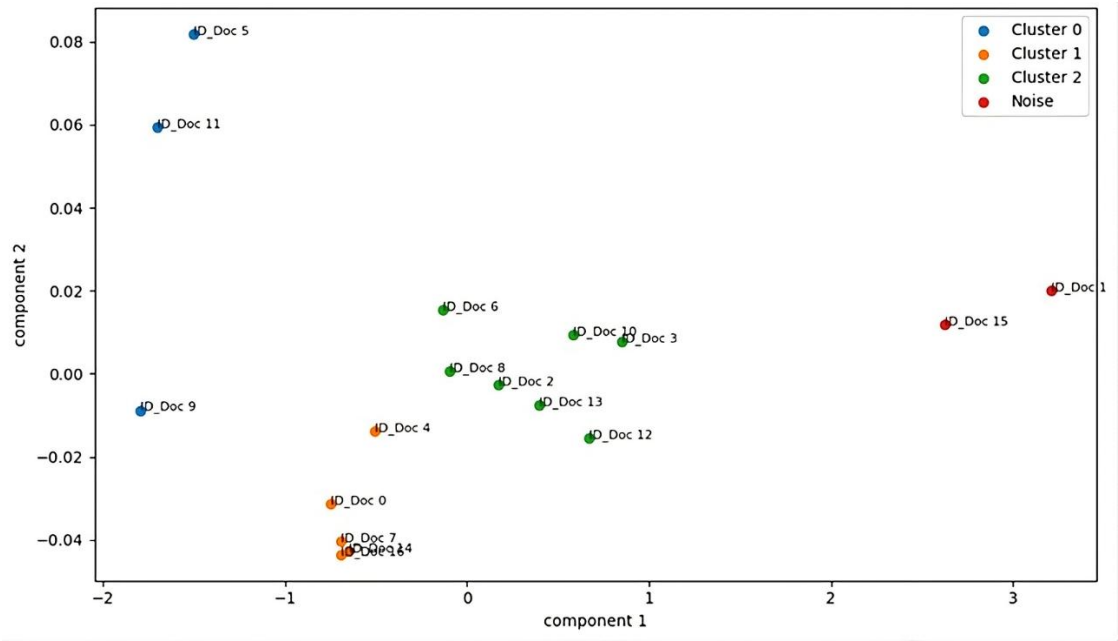
Jurnal Sarjana Teknik Informatika
Vol. 13., No. 3, Oktober 2025, pp. 120-128

Tabel 1 Hasil Pengelompokan Dokumen

Cluster	Jumlah
Cluster 0	3
Cluster 1	5
Cluster 2	7
Noise	2

Berdasarkan hasil pengelompokan, dokumen-dokumen tersebut dibagi menjadi empat kelompok. Cluster 2 menjadi grup terbesar dengan total 7 dokumen. Ini menunjukkan bahwa sebagian besar dokumen memiliki karakteristik yang paling mirip satu sama lain. Selanjutnya Cluster 1 terdiri dari 5 dokumen, sedangkan Cluster 0 berisi 3 dokumen. Kedua *cluster* ini mewakili kelompok yang lebih kecil dengan karakteristik internal yang serupa, tetapi berbeda dari cluster lainnya. Selain itu, ada 2 dokumen yang dikategorikan sebagai *Noise*. Artinya, dokumen tersebut merupakan pencilan atau anomali yang tidak memiliki cukup kesamaan dengan grup mana pun, sehingga tidak dapat diklasifikasikan ke dalam cluster mana pun.

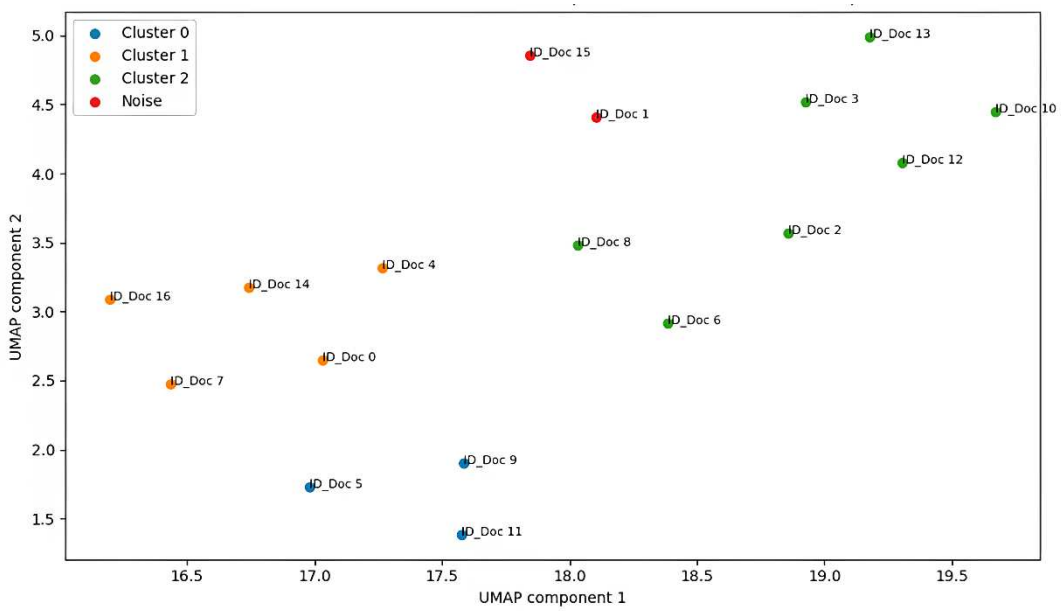
Evaluasi kinerja klaster dilakukan dengan menggunakan skor Silhouette yang menghasilkan nilai sebesar 0,6653. Skor ini, meskipun menunjukkan pemisahan *cluster* yang moderat, masih dapat diterima mengingat karakteristik data teks akademik yang penuh dengan '*noise*', di mana semantik dan parafrase sering tumpang tindih. Nilai ini ditemukan dengan melihat seberapa dekat suatu objek dengan objek lain dalam kelompoknya (disebut kohesi) dan seberapa jauh jaraknya dari objek dalam kelompok lain (disebut pemisahan). Kemudian, hasilnya dirata-ratakan untuk seluruh objek, memberikan skor yang berkisar dari -1 hingga 1. Skor mendekati 1 berarti cluster sangat rapat dan terpisah dengan jelas, sedangkan skor mendekati 0 berarti cluster tumpang tindih dan tidak terpisah dengan baik.



Gambar 2 Visualisasi Cluster Dokumen (Doc2Vec dan HDBSCAN)

Dibandingkan dengan metode lain seperti K-Means, kinerja ini terbukti jauh lebih unggul [16]. Dalam studi lain [20], pengelompokan K-Means dalam dokumen serupa hanya menghasilkan koefisien silhouette yang sangat rendah, berkisar antara 0,0027 hingga 0,0035. Perbedaan signifikan ini menegaskan bahwa kombinasi Doc2Vec dan HDBSCAN jauh lebih efektif dalam mengelompokkan dokumen secara bermakna, bahkan dalam kondisi data yang menantang. Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa pengelompokan dalam teks akademik seringkali menghasilkan nilai Silhouette yang tidak terlalu tinggi, namun tetap berhasil menciptakan pengelompokan yang relevan.

Selain evaluasi numerik, visualisasi hasil pengelompokan juga dihasilkan. Gambar 2 menyajikan proyeksi dua dimensi dari penyematan dokumen yang dikelompokkan oleh HDBSCAN, sedangkan Gambar 3 menunjukkan kluster yang sama setelah menerapkan UMAP untuk pengurangan dimensi.



Gambar 3 Visualisasi Cluster Dokumen (Doc2Vec, HDBSCAN, dan UMAP)

Kedua gambar tersebut memberikan wawasan yang saling melengkapi. Pada Gambar 2, cluster dapat dibedakan dengan outlier yang teridentifikasi dengan jelas, sedangkan Gambar 3 menggunakan UMAP mendemonstrasikan pembentukan grup yang lebih kompak, menyoroti keunggulan teknik visualisasi dalam menafsirkan embeddings berdimensi tinggi. Penggunaan label dokumen yang dianonimkan memastikan kepatuhan terhadap praktik penelitian etis dengan melindungi privasi peserta.

Tabel 2 Contoh Dokumen Representatif per Cluster

ID Dokumen	Topik	Cluster
ID_Doc 0	Jaringan Syaraf Tiruan	Cluster 1
ID_Doc 1	Jaringan Syaraf Biologi	Noise
ID_Doc 2	Pelatihan Jaringan Syaraf Tiruan	Cluster 2
ID_Doc 3	Pelatihan Jaringan Syaraf Tiruan	Cluster 2
ID_Doc 4	Pelatihan Jaringan Syaraf Tiruan	Cluster 1
ID_Doc 5	CNN dan RNN	Cluster 0

Analisis dari Tabel 2 memperkuat temuan visual pada Gambar 2 dan 3. Meskipun visualisasi UMAP menunjukkan pembentukan klaster yang kompak, pembedahan isi dokumen secara manual membuktikan bahwa pengelompokan tersebut bukanlah kebetulan, melainkan didasarkan pada kesamaan topik substantif. Misalnya, dokumen yang membahas "Pelatihan Jaringan Syaraf Tiruan" (Cluster 2) berhasil dipisahkan dari dokumen yang lebih umum membahas "Jaringan Syaraf Tiruan" (Cluster 1), menunjukkan kemampuan model untuk menangkap nuansa topik.

Selain itu, keberhasilan model mengidentifikasi "Jaringan Syaraf Biologi" sebagai *noise* (Dokumen ID_Doc 1) menegaskan bahwa algoritma tidak hanya mengandalkan kemiripan kata kunci, tetapi juga memahami perbedaan kontekstual yang signifikan. Hal ini menunjukkan bahwa metode visualisasi embedding mampu menjadi alat diagnostik yang efektif untuk memahami mengapa sebuah cluster terbentuk, sehingga memungkinkan analisis yang lebih mendalam dan bukan sekadar deskripsi visual semata.

Dari pengelompokan tersebut dapat dilakukan beberapa pengamatan. Pertama, pengelompokan besar di Cluster 2 menunjukkan potensi redundansi atau plagiarisme, karena banyak dokumen

memiliki kesamaan semantik yang tinggi. Cluster 1 dan Cluster 0 mewakili kesamaan yang lebih kecil namun berbeda, yang mungkin menunjukkan kasus plagiarisme lokal atau referensi bersama. Dokumen Kebisingan, di sisi lain, kemungkinan besar mencerminkan orisinalitas asli atau perbedaan dalam gaya penulisan dan konten. Skor Silhouette menunjukkan seberapa baik pengelompokan tersebut, dan visualisasi UMAP memberikan bukti yang jelas bahwa kelompok-kelompok tersebut terpisah dengan baik. Bersama-sama, hasil ini memvalidasi bahwa alur Doc2Vec-HDBSCAN efektif untuk deteksi plagiarisme, karena menyeimbangkan pengelompokan semantik dengan identifikasi anomali. Pendekatan ini menawarkan alat praktis bagi pendidik untuk memantau integritas akademik di lingkungan pembelajaran digital.

Meskipun penelitian ini menunjukkan hasil yang potensial, terdapat beberapa keterbatasan yang dapat menjadi fokus penelitian lanjutan. Pertama, penggunaan Doc2Vec sebagai model *embedding* masih memiliki batasan dalam menangkap makna ganda (*polysemy*) dari sebuah kata. Kinerja model ini juga sangat bergantung pada ukuran dan kualitas korpus data pelatihan. Kedua, HDBSCAN meskipun kuat, memerlukan penyetelan parameter yang cermat, seperti *min_cluster_size* dan *min_samples*, untuk mendapatkan hasil optimal. Keterbatasan lain adalah kurangnya perbandingan langsung dengan model *embedding* yang lebih modern seperti BERT atau GPT, yang dikenal mampu menangkap representasi kontekstual yang jauh lebih kaya. Oleh karena itu, arah penelitian selanjutnya dapat mencakup eksplorasi penggunaan BERT untuk menghasilkan *document embeddings* dan membandingkan performanya dengan Doc2Vec. Selain itu, pengujian pada korpus multi-bahasa atau dokumen dengan format yang lebih kompleks juga bisa menjadi area pengembangan di masa depan.

4. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa integrasi metode Doc2Vec dan HDBSCAN efektif untuk deteksi plagiarisme semantik. Dengan mengelompokkan dokumen berdasarkan kemiripan makna, pendekatan ini secara akurat mengidentifikasi dokumen yang terindikasi plagiat, yang ditunjukkan oleh visualisasi UMAP dan skor Silhouette sebesar 0,6653. Kontribusi utama penelitian ini adalah menawarkan kerangka kerja *machine learning* yang lebih akurat, melampaui metode pencocokan kata tradisional, dan menyediakan alat praktis bagi pendidik untuk menjaga integritas akademik. Meskipun efektif, penelitian ini memiliki keterbatasan karena diuji pada kumpulan data yang relatif kecil, yang membatasi generalisasi temuan. Oleh karena itu, penelitian di masa depan disarankan untuk mengeksplorasi penggunaan *embedding* yang lebih modern seperti BERT, memvalidasi model pada kumpulan data yang lebih besar, dan mengembangkan pendekatan hibrid untuk meningkatkan akurasi deteksi.

Deklarasi

Kontribusi Penulis. Semua penulis berkontrbusi secara bersama-sama dengan kontributor utama dalam artikel ini. Semua penulis membaca dan menyetujui versi akhir dari artikel yang diajukan.
Pernyataan Pendanaan. Tidak ada penulis yang menerima dana atau hibah dari lembaga atau badan pendanaan untuk penelitian ini.
Konflik Kepentingan. Penulis menyatakan tidak ada konflik kepentingan.
Informasi Tambahan. Tidak ada informasi tambahan dalam artikel ini.

Daftar Pustaka

[1] R. P. Pratama, M. Faisal, and A. Hanani, "Deteksi plagiarisme pada dokumen jurnal menggunakan metode cosine similarity", *SMARTICS*, vol. 5, no. 1, pp. 22–26, Apr. 2019, doi: 10.21067/smartics.v5i1.2848.

[2] E. A. Septiani and I. Arfiani, "Pengenalan makanan khas daerah riau dengan augmented reality berbasis android", *Jurnal Sarjana Teknik Informatika*, vol. 12, no. 3, pp. 111–120, Oct. 2024, doi: 10.12928/jstie.v12i3.30009.

[3] B. H. Mutongoza and B. E. Olawale, “Safeguarding academic integrity in the face of emergency remote teaching and learning in developing countries,” *Perspectives in Education*, vol. 40, no. 1, pp. 234–249, Mar. 2022, doi: 10.18820/2519593X/pie.v40.i1.14.

[4] M. A. Pratiwi and N. Aisya, “Fenomena plagiarisme akademik di era digital,” *Publishing Letters*, vol. 1, no. 2, pp. 16–33, Jul. 2021, doi: 10.48078/publetters.v1i2.23.

[5] H. Jolanda Pentury, I. Bolo Rangka, and A. Dewi Anggraeni, “Peningkatan kemampuan pedagogik guru dalam pembelajaran daring melalui penerapan kuis interaktif daring,” *Jurnal Surya Masyarakat*, vol. 3, no. 2, May 2021, doi: 10.26714/jsm.3.1.2020.109-114.

[6] Pupovac Vanja, “The frequency of plagiarism identified by text-matching software in scientific articles: a systematic review and meta-analysis,” *Scientometrics*, vol. 126, no. 11, pp. 8981–9003, Sep. 2021, doi: 10.1007/s11192-021-04140-5.

[7] S. Sunardi, A. Yudhana, and I. A. Mukaromah, "Implementasi deteksi plagiarisme menggunakan metode n-gram dan jaccard similarity terhadap algoritma winnowing," *Transmisi: Jurnal Ilmiah Teknik Elektro*, vol. 20, no. 3, pp. 105-110, Okt. 2018, doi: 10.14710/transmisi.20.3.105-110.

[8] M. F. Manzoor, M. S. Farooq, M. Haseeb, U. Farooq, S. Khalid, and A. Abid, “Exploring the landscape of intrinsic plagiarism detection: benchmarks, techniques, evolution, and challenges,” *IEEE Access*, vol. 11, pp. 140519–140545, Nov. 2023, doi: 10.1109/ACCESS.2023.3338855.

[9] D. A. Suryaningrum, R. Syaifudin, and H. R. P. Putra, “Integrasi word embeddings dan inverse book frequency dalam pembobotan term untuk peningkatan pencarian dokumen,” *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 9, no. 4, pp. 2529–2537, Dec. 2024, doi: 10.29100/jipi.v9i4.7557.

[10] K. I. Gunawan and J. Santoso, “Multilabel text classification menggunakan svm dan doc2vec classification pada dokumen berita bahasa indonesia,” *Journal of Information System,Graphics, Hospitality and Technology*, vol. 3, no. 01, pp. 29–38, Apr. 2021, doi: 10.37823/insight.v3i01.126.

[11] N. A. Wahyuni, M. N. Hayati, D. Nanda, and A. Rizki, “Metode hierarchical density-based spatial clustering of application with noise (hdbscan) pada wilayah desa/kelurahan tertinggal di kabupaten kutai kartanegara (studi kasus : data hasil pendataan potensi desa (podes) tahun 2018),” *Jurnal EKSPONENSIAL*, vol. 12, no. 1, 2021, doi: 10.30872/eksponensial.v12i1.758.

[12] A. Nurmasani and Y. Pristyanto, “Algoritme stacking untuk klasifikasi penyakit jantung pada dataset imbalanced class,” *Jurnal Pseudocode*, vol. 1, Feb. 2021, doi: 10.33369/pseudocode.8.1.21-26.

[13] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, “Komparasi ekstraksi fitur dalam klasifikasi teks multilabel menggunakan algoritma machine learning,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022, doi: 10.30812/matrik.v21i3.1851.

[14] Y. M. Pranoto, A. N. Handayani, H. W. Herwanto, and Y. Kristian, “Optimizing product matching in e-commerce with doc2vec: leveraging hierarchical clustering parameters based on product titles,” *ECTI Transactions on Computer and Information Technology*, vol. 18, no. 3, pp. 396–405, Jul. 2024, doi: 10.37936/ecti-cit.2024183.256164.

[15] R. González-Alemán *et al.*, “MDSCAN: RMSD-based HDBSCAN clustering of long molecular dynamics,” *Bioinformatics*, vol. 38, no. 23, pp. 5191–5198, Dec. 2022, doi: 10.1093/bioinformatics/btac666.

[16] G. Stewart, and M. Al-Khassaweneh, “An implementation of the hdbscan* clustering algorithm”, *Applied Sciences*, vol. 12, no. 5, pp. 2405, Feb. 2022, doi: 10.3390/app12052405.

[17] O. Purwaningrum, Y. Y. Putra, and A. A. Arifiyanti, “Penentuan kelompok status gizi balita dengan menggunakan metode k-means,” *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 15, no. 2, 2021, doi: 10.32815/jitika.v15i2.594.

[18] M. Shutaywi and N. N. Kachouie, “Silhouette analysis for performance evaluation in machine learning with applications to clustering,” *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.

[19] I. Tri, A. Mawarni, and A. P. Wibawa, “Analisis persepsi pengguna dengan melihat kualitas layanan pada aplikasi mobile transportasi online,” *Inovbiz*, vol. 8, no. 1, pp. 23-28, Apr. 2020, doi: 10.35314/inovbiz.v8i1.1286.

[20] D. Yudhistira Wijaya Koesuma, A. Hernawan, and R. S. Bianco Huwae, "Document clustering terkait health news pada twitter data set menggunakan k-means clustering," *Eprints Universitas Mataram*, Jun. 2023, [Online]. Available: http://eprints.unram.ac.id/id/eprint/39758