

**IMPLEMENTASI *DEEP LEARNING* MENGGUNAKAN
HYBRID SENTENCE-TRANSFORMERS DAN *K-MEANS*
UNTUK PERBANDINGAN JURNAL**

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Mendapat Gelar Sarjana
Komputer (S.Kom) Program Studi Informatika



MUHAMMAD ASYGAR FAERUDDIN

105841104721

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MAKASSAR

2025

**IMPLEMENTASI *DEEP LEARNING* MENGGUNAKAN
HYBRID SENTENCE-TRANSFORMERS DAN *K-MEANS*
UNTUK PERBANDINGAN JURNAL**

**Diajukan sebagai Salah Satu Syarat untuk Mendapatkan
Gelar Sarjana Komputer (S.Kom) Program Studi Informatika**

Disusun Dan Diajukan Oleh :

MUHAMMAD ASYGAR FAERUDDIN

105841104721

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MAKASSAR

2025



MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
FAKULTAS TEKNIK



PENGESAHAN

Skripsi atas nama Muhammad Asygar Faeruddin dengan nomor induk Mahasiswa 105841104721, dinyatakan diterima dan disahkan oleh Panitia Ujian Tugas Akhir/Skripsi sesuai dengan Surat Keputusan Dekan Fakultas Teknik Universitas Muhammadiyah Makassar Nomor : 0004/SK-Y/55202/091004/2025, sebagai salah satu syarat guna memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar pada hari Sabtu, 30 Agustus 2025.

Panitia Ujian :

1. Pengawas Umum

Makassar,

15 Rabi'ul Awwal 1447 H

08 September 2025 M

a. Rektor Universitas Muhammadiyah Makassar

Dr. Ir. H. Abd. Rakhim Nanda, ST, MT, IPU

b. Dekan Fakultas Teknik Universitas Muhammadiyah Makassar

Prof. Dr. Eng. Muhammad Isran Ramli, ST, MT, ASEAN, Eng

2. Penguji

a. Ketua

: Dr. Ir. Zahir Zainuddin, M. Sc.

b. Sekertaris

: Desi Anggreani, S. Kom., M. T.

3. Anggota

1. Darniati, S. Kom., M. T.

2. Muhyiddin A. M. Hayat, S. Kom., M. T.

3. Chyquitha Danuputri, S. Kom., M. Kom.

Mengetahui :

Pembimbing I

Pembimbing II

Ir. Muhammad Faisal, S. Si., M. T., Ph. D., IPM.

Rizki Yuliana Bakti, S. T., M. T.

Dekan



Ir. Muh. Syafiq S. Kuba, S. T., M. T.

NPM. 795 288

Gedung Menara Iqra Lantai 3

Jl. Sultan Alaaddin No. 259 Telp. (0411) 866 972 Fax (0411) 865 588 Makassar 90221

Web: <https://teknik.unismuh.ac.id/>, e-mail: teknik@unismuh.ac.id





HALAMAN PENGESAHAN

Tugas Akhir ini diajukan untuk memenuhi syarat ujian guna memperoleh gelar Sarjana Komputer (S.Kom) Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.

Judul Skripsi : **IMPLEMENTASI *DEEP LEARNING* MENGGUNAKAN *HYBRID SENTENCE-TRANSFORMERS* DAN *K-MEANS* UNTUK PERBANDINGAN JURNAL**

Nama : Muhammad Asygar Faeruddin

Stambuk : 105 84 11047 21

Makassar, 08 September 2025

Telah Diperiksa dan Disetujui
Oleh Dosen Pembimbing;

Pembimbing I

Pembimbing II

Ir. Muhammad Faisal, S.Si., M.T., Ph.D., IPM.

Rizki Yusliana Bakti, S.T., M.T.

Mengetahui,

Ketua Prodi Informatika



Rizki Yusliana Bakti, S.T., M.T.

NBM : 1307 284



MOTTO DAN PERSEMBAHAN

Motto

“Lakukan yang mustahil, lihat yang tak terlihat! Tendang rasionalitas dan lakukan yang tak mungkin!”

Persambahan

Bismillahirrahmanirrahim. Skripsi ini saya persembahkan dengan penuh rasa syukur kepada Allah SWT, yang dengan limpahan rahmat, hidayah, dan karunia-Nya sehingga saya dapat menyelesaikan skripsi ini. Dengan penuh kerendahan hati, skripsi ini kupersembahkan kepada ayahanda tercinta Faeruddin dan ibunda tersayang Nurjana Sasahi, yang selalu menjadi sumber doa, kasih sayang, serta semangat dalam setiap langkah perjuangan. Terima kasih yang tak terhingga atas pengorbanan, cinta, dan keikhlasan yang telah diberikan, sehingga saya mampu berdiri hingga tahap ini. Skripsi ini juga kupersembahkan untuk diriku sendiri, sebagai bukti bahwa usaha, kesabaran, dan keyakinan akan selalu berbuah hasil indah di penghujung perjalanan. Ucapan terima kasih saya sampaikan kepada seluruh dosen dan pembimbing di Program Studi Informatika, khususnya Fakultas Teknik Universitas Muhammadiyah Makassar, atas ilmu, bimbingan, dan inspirasi yang telah diberikan. Tidak lupa kepada seluruh civitas akademika Fakultas Teknik, yang telah menjadi bagian penting dalam proses belajar dan pengalaman berharga selama menempuh pendidikan. Skripsi ini juga kupersembahkan kepada teman-teman angkatan 2021 Fakultas Teknik, khususnya Program Studi Informatika, yang telah menjadi sahabat seperjuangan, berbagi suka duka, tawa, serta semangat dalam menapaki jalan panjang perkuliahan. Semoga Skripsi ini dapat memberi manfaat, menjadi langkah kecil menuju perubahan yang lebih baik, dan menjadi persembahan tulus bagi semua pihak yang telah hadir dan mendukung perjalanan ini.

ABSTRAK

MUHAMMAD ASYGAR FAERUDDIN. Implementasi *Deep Learning* Menggunakan *Hybrid Sentence-Transformers* Dan *K-Means* Untuk Perbandingan Jurnal (Dibimbing Oleh Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM., dan Rizki Yusliana Bakti, S.T., M.T.,).

Pertumbuhan eksponensial publikasi ilmiah menyulitkan proses perbandingan jurnal secara manual. Penelitian ini mengimplementasikan sistem otomatis berbasis *deep learning* yang menggabungkan *Sentence-Transformers* dan *K-Means Clustering* untuk analisis kemiripan semantik. Metode penelitian meliputi perhitungan skor *cosine similarity* dari abstrak jurnal yang kemudian diklusterisasi menggunakan *K-Means*. Hasil evaluasi dengan *Elbow Method*, *Silhouette Score*, dan *DBI* menetapkan empat kategori kemiripan sebagai jumlah kluster optimal: Tidak Relevan, Sedikit Berkaitan, Cukup Berkaitan, dan Sangat Berkaitan. Sistem yang diimplementasikan mampu membandingkan dua jurnal dengan waktu komputasi rata-rata 15,25 detik, serta menyajikan skor kemiripan beserta klasifikasi hasilnya. Pengujian fungsional membuktikan sistem ini efektif dan efisien dalam memfasilitasi analisis komparatif jurnal, secara signifikan mempercepat proses tinjauan literatur bagi peneliti.

Kata Kunci: Perbandingan Jurnal, *Deep Learning*, *Sentence-Transformers*, *K-Means Clustering*, Kemiripan Semantik, *Natural Language Processing*.

ABSTRACT

MUHAMMAD ASYGAR FAERUDDIN. *Implementation of Deep Learning Using Hybrid Sentence-Transformers and K-Means for Journal Comparison (Supervised by Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM., and Rizki Yusliana Bakti, S.T., M.T.,).*

The exponential growth of scientific publications challenges manual journal comparison. This research implements an automated deep learning-based system combining Sentence-Transformers and K-Means Clustering for semantic similarity analysis. The method involves calculating cosine similarity scores from journal abstracts, which are then clustered using K-Means. Evaluation using the Elbow Method, Silhouette Score, and DBI established four optimal similarity categories: Not Relevant, Slightly Related, Moderately Related, and Highly Related. The implemented system can compare two journals with an average computation time of 15.25 seconds, presenting the similarity score and its resulting classification. Functional testing confirms the system is effective and efficient for comparative journal analysis, significantly accelerating the literature review process for researchers.

Keywords: *Journal Comparison, Deep Learning, Sentence-Transformers, K-Means Clustering, Semantic Similarity, Natural Language Processing.*

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Segala puji bagi Allah Subhanahu Wata'ala, atas limpahan rahmat dan Karunia-Nya sehingga penulis dapat menyelesaikan penyusunan Laporan Skripsi ini yang berjudul **“IMPLEMENTASI DEEP LEARNING MENGGUNAKAN HYBRID SENTENCE-TRANSFORMERS DAN K-MEANS UNTUK PERBANDINGAN JURNAL”**. Shalawat beserta salam senantiasa penulis panjatkan kepada Nabi besar Muhammad SAW sebagai uswatun hasanah dan rahmatan lil alamin.

Dalam penyusunan Skripsi ini penulis mengucapkan banyak terima kasih kepada semua pihak yang telah membantu dan memberikan motivasi serta bimbingan dan arahan dalam penyusunan Skripsi ini, terutama kepada:

1. Kedua orang tua yang selalu memberikan dukungan baik berupa moral, materi, dan spiritual agar dapat terselesaikannya penyusunan skripsi ini.
2. Bapak Dekan **Ir. Muhammad Syafaat S. Kuba, S.T., M.T.**, selaku Dekan Fakultas Teknik Universitas Muhammadiyah Makassar.
3. Ibu **Rizki Yusliana Bakti S.T., M.T.**, selaku Ketua Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar.
4. Bapak **Ir. Muhammad Faisal, S.SI., M.T., Ph.D., IPM** selaku Dosen Pembimbing 1 sekaligus sebagai Dosen Pembimbing Akademik yang telah memberikan arahan dan bimbingan serta saran yang sangat berarti dalam penyusunan skripsi ini.
5. Serta Ibu **Rizki Yusliana Bakti S.T., M.T.**, juga selaku Dosen Pembimbing 2 yang telah memberikan arahan dan bimbingan serta saran yang sangat berarti dalam penyusunan skripsi ini.
6. Seluruh Dosen Fakultas Teknik Program Studi Informatika Universitas Muhammadiyah Makassar yang telah memberikan ilmu dan bantuannya dalam penulisan skripsi ini.

7. Teman-teman angkatan 2021 Program Studi Informatika Fakultas Teknik Universitas Muhammadiyah Makassar, khususnya Teman-teman Kelas B Informatika, terima kasih atas dukungan dan doanya.
8. Kepada semua pihak yang tidak bisa penulis sebutkan satu persatu, penulis mengucapkan banyak terima kasih yang sebesar-besarnya.

Demikian Skripsi ini disusun dengan penuh kesadaran bahwa masih terdapat kekurangan yang perlu diperbaiki di masa mendatang. Penulis berharap karya ini dapat memberikan manfaat dan menjadi pijakan untuk penelitian selanjutnya.

Billahi fii sabililhaq, fastabiqul khairat.

Wassalamualaikum Wr. Wb.

Makassar, 30 Agustus 2025

Penulis

Muhammad Asygar Faeruddin



DAFTAR ISI

ABSTRAK	ii
KATA PENGANTAR.....	iv
DAFTAR ISI.....	vi
DAFTAR GAMBAR.....	viii
DAFTAR TABEL.....	ix
DAFTAR LAMPIRAN	x
DAFTAR ISTILAH	xi
BAB I PENDAHULUAN.....	1
A. Latar Belakang.....	1
B. Rumusaan Masalah.....	3
C. Tujuan Peniltian.....	3
D. Manfaat Peniltian.....	3
E. Ruang Lingkup Peniltian.....	4
F. Sistematika Penulisan.....	5
BAB II TINJAUAN PUSTAKA.....	6
A. Landasan Teori	6
B. Peniltian Terkait	13
C. Kerangka Berpikir	16
BAB III METODE PENELITIAN	17
A. Tempat dan Waktu Peniltian	17
B. Alat dan Bahan	17
C. Perancangan Sistem.....	17
D. Teknik Pengujian Sistem.....	22
E. Teknik Analisis Data	22
BAB IV HASIL DAN PEMBAHASAN.....	24
A. Dataset Jurnal	24
B. <i>Pseudo-Labeling</i> Dataset Jurnal	25
C. Evaluasi Klaster Optimal.....	26

D. Proses Klastering <i>Pseudo-Label</i>	30
E. Implementasi Sistem.....	33
F. Pengujian Sistem	38
BAB V KESIMPULAN DAN SARAN	42
A. Kesimpulan.....	42
B. Saran	43
DAFTAR PUSTAKA	44
LAMPIRAN.....	48



DAFTAR GAMBAR

Gambar 1. Kerangka Berpikir.....	16
Gambar 2. <i>Flowchart</i> Perancangan Model dan <i>Clustering</i>	18
Gambar 3. <i>Flowchart</i> Pengolahan Jurnal dan Output.....	20
Gambar 4. Sampel Dataset Jurnal (19 Baris Pertama).....	24
Gambar 5. Distribusi Skor <i>Pseudo-Label</i> (<i>Histogram</i> & KDE).....	26
Gambar 6. Dataset sebelum proses <i>Pseudo-Labeling</i>	26
Gambar 7. Dataset sesudah proses <i>Pseudo-Labeling</i>	26
Gambar 8. Grafik <i>Elbow Method</i> dari <i>Pseudo-Label</i>	28
Gambar 9. Grafik <i>Silhouette Score</i> (Biru) dan DBI (Jingga) dari <i>Pseudo-Label</i> ..	28
Gambar 10. Visualisasi klastering <i>pseudo-label</i> pada masing-masing kategori....	32
Gambar 11. Tampilan halaman utama sistem (<i>home page</i>).....	34
Gambar 12. Proses unggah dua jurnal oleh pengguna.....	35
Gambar 13. Pratinjau file jurnal sebelum proses komparasi.....	35
Gambar 14. Proses komputasi sistem dalam melakukan perbandingan	36
Gambar 15. Hasil komparasi dan ringkasan jurnal (Bahasa Inggris).....	37
Gambar 16. Hasil komparasi dan ringkasan jurnal (Bahasa Indonesia)	37

DAFTAR TABEL

Tabel 1. Peniltian Terkait.....	13
Tabel 2. Nilai Metrik Evaluasi untuk Setiap Jumlah Klaster.....	29
Tabel 3. Perbandingan Evaluasi Jumlah Klaster.....	29
Tabel 4. Statistik Deskriptif <i>Pseudo-Label</i> per Klaster	32
Tabel 5. Hasil Pengujian Sistem dengan Metode <i>Blackbox Testing</i>	38
Tabel 6. Evaluasi Akurasi Fungsional dan Kinerja Sistem.....	40



DAFTAR LAMPIRAN

Lampiran 1. Dataset Abstrak Jurnal.....	48
Lampiran 2. Dataset Setelah <i>Pseudo-Labeling</i>	50
Lampiran 3. Dataset Setelah <i>Clustering</i>	52
Lampiran 4. Source Code.....	54
Lampiran 5. Surat Keterangan Bebas Plagiat	63
Lampiran 6. Uji Plagiat BAB 1-5	64



DAFTAR ISTILAH

API (Application Programming Interface)	Antarmuka pemrograman aplikasi yang memungkinkan berbagai aplikasi perangkat lunak untuk berkomunikasi satu sama lain.
Black-Box Testing	Metode pengujian perangkat lunak yang berfokus pada verifikasi fungsionalitas eksternal sistem tanpa menganalisis struktur internal kode, dengan mengamati hubungan antara input dan output.
Clustering	Salah satu teknik dalam <i>unsupervised learning</i> yang bertujuan untuk mengelompokkan data ke dalam beberapa kelompok (klaster) berdasarkan tingkat kemiripan antar data.
Cosine Similarity	Metrik yang digunakan untuk mengukur kesamaan semantik antara dua vektor dengan menghitung besar sudut di antara keduanya. Nilai yang mendekati 1 menunjukkan kemiripan yang sangat tinggi.
Centroid	Pusat dari sebuah klaster dalam algoritma <i>K-Means</i> . Posisinya diperbarui secara iteratif dengan menghitung rata-rata dari semua titik data dalam klaster tersebut.
Davies-Bouldin Index (DBI)	Metrik evaluasi klaster yang menilai rasio antara sebaran dalam klaster dengan jarak antar klaster. Nilai yang lebih kecil menunjukkan pengelompokan yang lebih baik
Deep Learning	Sub-bidang dari <i>machine learning</i> yang menggunakan jaringan saraf tiruan dengan banyak lapisan (<i>deep neural networks</i>) untuk mempelajari pola kompleks dari data dalam jumlah besar.
Elbow Method	Metode untuk menentukan jumlah klaster optimal dalam <i>K-Means</i> dengan melihat titik "siku" (<i>elbow</i>)

	pada grafik, di mana penurunan nilai <i>inertia</i> (WCSS) mulai melambat secara signifikan.
<i>Embedding</i>	Representasi numerik (vektor) dari data teks yang dapat diproses oleh model <i>machine learning</i> . <i>embedding</i> mengubah seluruh kalimat menjadi sebuah vektor yang menangkap makna semantiknya.
<i>K-Means</i>	Salah satu algoritma partitional <i>clustering</i> yang membagi data ke dalam sejumlah <i>k</i> kluster berdasarkan kedekatan setiap titik data dengan <i>centroid</i> kluster
<i>Pseudo-Labeling</i>	Teknik dalam pembelajaran <i>semi-supervised</i> di mana sebuah model yang telah dilatih pada data berlabel digunakan untuk memprediksi label pada data yang tidak berlabel.
<i>Semantic Similarity</i>	Pendekatan dalam NLP yang bertujuan untuk mengukur tingkat kemiripan makna antara dua unit teks (kalimat atau dokumen), bukan hanya berdasarkan pencocokan kata secara harfiah.
<i>Sentence-Transformers (SBERT)</i>	Modifikasi dari arsitektur BERT yang dirancang khusus untuk menghasilkan <i>sentence embeddings</i> yang bermakna secara semantik, sehingga efisien untuk tugas perbandingan kemiripan teks.
<i>Silhouette Score</i>	Metrik evaluasi kluster yang mengukur seberapa baik setiap objek cocok dengan klasternya sendiri dibandingkan dengan kluster lain yang terdekat. Nilai yang mendekati +1 menunjukkan pembentukan kluster yang baik.
<i>Transformer</i>	Arsitektur <i>deep learning</i> yang menggunakan mekanisme <i>self-attention</i> untuk menangani data sekuensial, memungkinkannya menangkap hubungan kontekstual secara global dalam teks.

***Unsupervised
Learning***

Pendekatan dalam *machine learning* di mana sistem dilatih pada data yang tidak memiliki label, dengan tujuan menemukan struktur atau pola tersembunyi secara mandiri.



BAB I

PENDAHULUAN

A. Latar Belakang

Pertumbuhan jumlah publikasi ilmiah di dunia mengalami peningkatan eksponensial, dengan rata-rata laju pertumbuhan sekitar 4,10% per tahun dan waktu penggandaan (*doubling time*) sebesar 17,3 tahun (Bornmann et al., 2021). Lonjakan ini, didorong oleh kemajuan teknologi informasi dan ketersediaan database digital, menimbulkan tantangan bagi peneliti dalam membandingkan, menelusuri, serta memahami keterkaitan antar jurnal yang relevan. Metode manual untuk menilai kemiripan antar dokumen seringkali memakan waktu lama dan rentan bias, sehingga diperlukan pendekatan otomatis yang mampu mengidentifikasi kemiripan semantik antar dokumen ilmiah secara efisien dan konsisten, sekaligus memfasilitasi perbandingan jurnal berdasarkan tingkat kemiripannya.

Teknologi pemrosesan bahasa alami (*Natural Language Processing*) kini banyak mengandalkan model berbasis *Transformer*, seperti BERT dan variannya, yang mampu menangkap hubungan kontekstual antar kata dengan lebih akurat. Berdasarkan arsitektur ini, *Sentence-Transformer* models dirancang khusus untuk tugas *semantic similarity*, menghasilkan *dense semantic embeddings* yang mempertahankan relasi makna teks dalam ruang vektor. Dengan representasi vektor ini, perhitungan kemiripan antar kalimat atau dokumen menggunakan metrik *cosine similarity* menjadi sangat andal, sehingga model ini sangat sesuai untuk tugas analisis kemiripan jurnal ilmiah berdasarkan kemiripan semantik (Colangelo et al., 2025).

Penggunaan *Sentence-Transformers* pada kombinasi dengan algoritma *K-Means Clustering* telah terbukti efektif untuk mengelompokkan dokumen berdasarkan kedekatan semantik. Sebagaimana penelitian yang dilakukan oleh Kusumaningrum et al. (2024), pada mengelompokkan opini masyarakat terhadap pelestarian ekosistem mangrove dan berhasil mengidentifikasi topik

secara lebih kontekstual dibandingkan pendekatan LDA konvensional. Hal ini menunjukkan bahwa *Sentence-BERT* mampu menangkap nuansa makna kalimat dan *K-Means* menjadi metode *clustering* yang efisien untuk dokumen tanpa label (*unsupervised*).

Penelitian terdahulu menunjukkan potensi dan batasan berbagai teknik *embedding* dan *clustering*. Simanjuntak et al. (2023) melaporkan *Word2Vec* dan *K-Means* efektif untuk klasterisasi berita Indonesia, namun hanya pada level kata tanpa konteks kalimat penuh. Demikian pula, Aszani et al. (2023) menerapkan TF-IDF dengan *Cosine Similarity* pada data rekam medis, tetapi kurang optimal menangkap relasi semantik antar kalimat. Selain itu, Subakti et al. (2022) membandingkan representasi teks menggunakan BERT dan TF-IDF dalam tugas *clustering*, dan menemukan bahwa BERT mengungguli TF-IDF pada 28 dari 36 metrik. Lebih lanjut, Ortakci et al. (2024) menginvestigasi berbagai varian model SBERT dan teknik *pooling*, dan menunjukkan bahwa *mean pooling* memberikan hasil *clustering* terbaik di sebagian besar tugas, menegaskan fleksibilitas SBERT dalam *transfer learning*. Prinsip *embeddings* berbasis *transformer* juga ditegaskan oleh Weng et al. (2022), yang memanfaatkan *embedding* semantik GPT-3 dan algoritma HDBSCAN untuk mengidentifikasi topik kunci dalam publikasi ilmiah. Oleh karena itu, penelitian ini mengusulkan penggunaan model *embedding* kontekstual berbasis *Sentence-Transformers* (SBERT) yang dikombinasikan dengan algoritma *K-Means*, dengan tujuan meningkatkan akurasi perbandingan kemiripan semantik antar jurnal ilmiah.

Meskipun beberapa penelitian sebelumnya telah menerapkan pendekatan semantik pada kumpulan teks, namun fokus penelitian tersebut masih terbatas di berbagai domain tertentu seperti media sosial, berita, dan data medis. Sedangkan pada proses analisis kemiripan semantik antar jurnal, masih terdapat keterbatasan ketersediaan dataset berlabel. Hal ini menyulitkan penerapan metode *supervised* yang memerlukan pelabelan manual dan *fine-tuning* model. Oleh karena itu, pendekatan tanpa pengawasan (*unsupervised*) seperti

kombinasi *Sentence-Transformers* dan *K-Means* menjadi solusi yang relevan untuk penilitian ini.

Berdasarkan permasalahan yang ditemukan, penelitian ini bertujuan menerapkan kombinasi *Sentence-Transformers* dan *K-Means* untuk mengelompokkan jurnal ilmiah berdasarkan kemiripan semantik abstraknya. Dalam rangka penerapan teori NLP dan *Machine Learning*, untuk membangun sistem otomatis yang dapat mempermudah analisis komparatif jurnal secara kontekstual.

B. Rumusaan Masalah

Berdasarkan konteks permasalahan yang telah dijelaskan di atas, permasalahan pokok yang dihadapi adalah:

1. Bagaimana penerapan *Sentence-Transformers* dan *K-Means* untuk membandingkan jurnal berdasarkan kemiripan semantik?
2. Apakah metode *Sentence-Transformers* dan *K-Means* efektif membantu peneliti dalam membandingkan jurnal secara semantik?

C. Tujuan Penilitian

Berdasarkan uraian rumusan masalah tersebut, maka dirumuskan tujuan penilitian tersebut:

1. Untuk menerapkan *Sentence-Transformers* dan *K-Means* dalam membandingkan jurnal berdasarkan kemiripan semantik.
2. Untuk menguji efektivitas metode dalam membantu peneliti membandingkan jurnal secara semantik.

D. Manfaat Penilitian

Adapun manfaat penilitian yang akan didapatkan dari penilitian ini yang berjudul Implementasi *Deep Learning* menggunakan *Hybrid Sentence-Transformers* dan *K-Means* untuk perbandingan jurnal :

1. Terhadap Penulis

Penelitian ini memberikan manfaat terhadap penulis untuk memperdalam pemahaman terhadap teknologi *Deep Learning*, khususnya pada penerapan model *Sentence-Transformers* dan *K-Means*. Dan juga penelitian ini menjadi sarana pembelajaran praktis dalam membangun sistem perbandingan jurnal berdasarkan kemiripan teks berbasis *Deep Learning*. Dengan begitu, penulis memperoleh pengalaman dalam merancang, mengimplementasikan, serta mengevaluasi sistem pembelajaran mesin.

2. Terhadap Peneliti atau Akademisi

Hasil penelitian ini dapat dimanfaatkan oleh Peneliti maupun Akademisi sebagai alternatif dalam menganalisis kemiripan antar jurnal secara semantik dan membandingkannya. Dengan adanya sistem yang mampu memberikan skor kemiripan dan label kategorikal peneliti dapat lebih memahami hubungan antar karya ilmiah yang sedang dikaji tanpa harus membaca secara manual isi jurnal.

E. Ruang Lingkup Penelitian

Dari analisis rumusan masalah di atas, dapat dirumuskan beberapa Batasan masalah sebagai berikut :

1. Penelitian ini berfokus pada penerapan *Sentence-Transformers* untuk perhitungan skor kemiripan dan *K-Means* untuk *Clustering* dalam analisis kemiripan jurnal ilmiah.
2. Dataset yang digunakan adalah jurnal dengan topik Teknologi, Kesehatan dan Ekonomi. Dataset di ambil melalui API dari *semantic scholar*.
3. Perbandingan jurnal dilakukan pada bagian abstrak kerana dianggap mewakili keseluruhan isi jurnal.
4. Hasil *similarity* dan *clustering* digunakan untuk menghasilkan deskripsi perbandingan dengan AI generatif sebagai nilai tambah sistem.

F. Sistematika Penulisan

Secara garis besar penulisan skripsi ini terbagi menjadi beberapa bab yang tersusun sebagai berikut:

BAB I PENDAHULUAN

Bab ini menerangkan secara singkat dan jelas mengenai latar belakang penulisan penelitian skripsi, rumusan masalah, tujuan dan manfaat, batasan permasalahan metodologi yang digunakan dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini membahas tentang teori-teori yang melandasi penulis dalam melaksanakan penelitian seperti metode dan algoritma yang terkait dengan penelitian ini. Pada bab ini juga berisi penjelasan dari penelitian sebelumnya yang relevan dengan penelitian ini.

BAB III METODE PENELITIAN

Pada bab ini membahas tentang metode yang digunakan dalam melakukan penelitian, waktu dan tempat penelitian, alat dan bahan penelitian, perancangan sistem penelitian dan teknik analisis data.

BAB II

TINJAUAN PUSTAKA

A. Landasan Teori

1. *Natural Language Processing (NLP)*

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan komputer untuk memahami, menginterpretasikan, dan menghasilkan bahasa manusia secara alami. Seiring berkembangnya teknologi, NLP menjadi komponen penting dalam berbagai aplikasi seperti *machine translation*, *sentiment analysis*, dan *information extraction*. NLP juga memungkinkan komputer untuk memproses bahasa manusia dengan bantuan kamus digital yang diperbarui secara berkelanjutan agar sesuai dengan perkembangan kosakata dan penggunaan bahasa sehari-hari. NLP++ menjadi salah satu *framework* yang efektif untuk meningkatkan akurasi proses NLP dalam sistem berskala besar, berkat kemampuan *multi-pass processing* dan integrasi *knowledge base* berbasis hierarki konseptual (Chikkarangiah et al., 2024).

Perkembangan NLP modern telah mengalami transformasi signifikan dari pendekatan statistik konvensional menuju paradigma berbasis *deep learning*. Pergeseran ini didorong oleh kemampuan jaringan saraf dalam menangkap pola semantik kompleks secara otomatis dari data alami, seperti yang ditunjukkan melalui adopsi arsitektur RNN, CNN, dan *Transformer* untuk tugas analisis sentimen, terjemahan mesin, dan pengenalan entitas. Model berbasis *Transformer* seperti BERT memperkenalkan mekanisme *self-attention* yang memungkinkan pemrosesan konteks secara *bidirectional*, mengatasi keterbatasan metode statistik dalam memahami hubungan antar kata jarak jauh. Integrasi teknik *hybrid* seperti kombinasi BiLSTM-CRF dan BERT serta pendekatan *ensemble* seperti LLM-Blender semakin memperkuat generalisasi sistem dengan menggabungkan keunggulan berbagai model (Jia et al., 2023).

Selain itu, teknik *deep learning* telah banyak diterapkan dalam bidang *computer vision*, seperti pengenalan objek berbasis gambar, dan konsep serupa dapat diterapkan dalam pemrosesan teks. Sistem berbasis *deep learning* memiliki kemampuan mengenali pola yang kompleks melalui representasi vektor kontekstual, yang bermanfaat untuk tugas klasifikasi dan analisis semantik pada teks alami (Octavian Ery Pamungkas et al., 2022).

2. *Transformers* dan *Sentence-Transformers*

Transformers merupakan arsitektur *deep learning* yang diperkenalkan untuk menangani data sekuensial dengan pendekatan *self-attention*. Keunggulan utamanya terletak pada kemampuannya untuk menangkap hubungan kontekstual secara global, baik dalam data teks maupun visual. Pada sebuah studi, arsitektur *transformer* terbukti efektif dalam melakukan prediksi *time series*, karena mampu menghindari masalah *vanishing gradient* serta mempertahankan informasi historis secara efisien. Model *transformer* menghasilkan nilai MAE dan RMSE yang lebih rendah dibanding pendekatan lainnya, menegaskan keunggulannya dalam konteks peramalan data yang kompleks (Kurnianingsih et al., 2023).

Pengembangan lebih lanjut dari arsitektur *Transformer* adalah *Sentence-Transformers* (SBERT) yang dikembangkan oleh Reimers dan Gurevych. SBERT memodifikasi struktur dasar BERT dengan menambahkan lapisan *pooling* setelah proses *encoding* untuk menghasilkan representasi vektor kalimat berdimensi tetap (*sentence embeddings*). Model ini dirancang untuk mengukur kemiripan semantik antar kalimat secara efisien, mendukung berbagai tugas seperti klasifikasi teks, *clustering*, dan *Semantic Textual Similarity* (STS). penggunaan SBERT sebagai metode *sentence embedding* terbukti mampu meningkatkan performa deteksi rumor di media sosial. Hasil eksperimen menunjukkan bahwa model berbasis SBERT menghasilkan peningkatan akurasi dan F1-score rata-rata sebesar 10% dibandingkan pendekatan berbasis *feature engineering* konvensional. Temuan ini menegaskan keunggulan SBERT dalam menangkap makna

kontekstual teks, sekaligus memperkuat relevansinya untuk diterapkan dalam analisis kemiripan semantik antar jurnal ilmiah (Anggrainingsih et al., 2024).

3. *Clustering*

Clustering merupakan salah satu teknik dalam *unsupervised learning* yang bertujuan untuk mengelompokkan data ke dalam beberapa kelompok (*cluster*) berdasarkan tingkat kemiripan antar data. Proses ini tidak membutuhkan data label seperti pada *supervised learning*, sehingga sangat efektif untuk eksplorasi data dan pemahaman pola tersembunyi. Setiap *cluster* terdiri atas data yang memiliki karakteristik serupa satu sama lain dan berbeda secara signifikan dari data dalam *cluster* lainnya. Metode ini telah digunakan secara luas dalam berbagai bidang seperti pengenalan pola, penambangan data, dan pengolahan citra (Zainuddin dan Risal, 2024).

Tujuan utama dari *clustering* adalah memaksimalkan homogenitas dalam satu *cluster* dan meminimalkan kesamaan antar *cluster*. Teknik ini memungkinkan analisis untuk memahami struktur alami dalam dataset yang kompleks. Dalam konteks industri berskala besar, *clustering* berbasis *K-Means* digunakan untuk mengelompokkan pekerja berdasarkan lokasi geografis guna membangun sistem keamanan yang efisien. Studi oleh Seo et al. (2022) mendemonstrasikan implementasi *clustering* untuk mengoptimalkan notifikasi ancaman bahaya (*threat notification*) melalui jaringan sensor nirkabel industri (IWSNs), di mana pengelompokan pekerja memungkinkan pembentukan *multicasting tree* dan *local flooding* yang mengurangi konsumsi energi hingga 36% dibanding metode konvensional. Sementara dalam konteks pertanian, *clustering* juga membantu pengelompokan hasil produksi petani untuk mendukung pengambilan keputusan berbasis data (Harahap et al., 2022).

Berbagai metode *clustering* telah dikembangkan, seperti *hierarchical clustering*, *density-based clustering* (DBSCAN), dan *partitional clustering* seperti *K-Means*. Pemilihan metode *clustering* yang

tepat bergantung pada jenis dan karakteristik data yang dianalisis. Kelebihan dari metode *clustering* adalah fleksibilitasnya dalam menangani data tak berlabel, namun kelemahannya terletak pada sensitivitas terhadap parameter awal seperti jumlah *cluster* dan pemilihan *centroid* (Annisa et al., 2020).

4. *K-Means*

K-Means merupakan salah satu algoritma *partitional clustering* yang paling banyak digunakan karena efisiensinya dalam membagi data ke dalam kelompok-kelompok berdasarkan kemiripan. Algoritma ini bekerja dengan cara menentukan jumlah *cluster* (k) yang diinginkan, kemudian memilih *centroid* awal secara acak. Setiap data akan dihitung jaraknya ke masing-masing *centroid* dan ditempatkan pada *cluster* terdekat. Perhitungan jarak ini umumnya menggunakan rumus *Euclidean Distance* (1):

$$d(O_i, O_j) = \sqrt{\sum_{d=1}^p (O_{id} - O_{jd})^2} \quad (1)$$

Di mana :

- O_i, O_j = objek data ke-I dan ke-j
- P = jumlah dimensi atau fitur
- $O_{id} - O_{jd}$ = nilai fitur ke-d dari objek O_i, O_j
- $d(O_i, O_j)$ = jarak antara objek O_i dan O_j

Setelah proses klasifikasi seluruh data selesai, posisi *centroid* untuk setiap *cluster* diperbarui melalui perhitungan rata-rata vektor anggota *cluster*. Prosedur pembaruan ini secara matematis dirumuskan pada Rumus 2:

$$C_i = \frac{1}{m_i} \sum_{x \in C_i} x \quad (2)$$

Di mana :

- C_i = *centroid* dari *cluster* ke-i
- m_i = jumlah data dalam *cluster* C_i

- $x \in C_i$ = setiap titik data dalam *cluster* C_i
- $\sum x$ = penjumlahan semua vektor data dalam *cluster*

Proses ini berulang hingga posisi *centroid* tidak berubah lagi atau hingga mencapai batas iterasi tertentu (Yee et al., 2023).

Penerapan *K-Means* terbukti efektif dalam berbagai domain. Misalnya, dalam penelitian deteksi stroke iskemik, kombinasi CNN dan Kernel *K-Means* berhasil mencapai akurasi 99% dalam klasifikasi citra MRI (Rustam et al., 2023). Pada bidang ekonomi, integrasi *K-Means* dengan *Multiple Linear Regression* mampu meningkatkan akurasi prediksi pendapatan rumah tangga dibanding model regresi biasa (Yee et al., 2023). Dalam pengelolaan data perangkat lunak, pengembangan metode *Point Center* untuk inisialisasi *centroid* mampu mengurangi error rata-rata hingga 12.82% dibanding metode acak konvensional (Annisa et al., 2020).

5. *Unsupervised Learning*

Unsupervised Learning atau pembelajaran tanpa pengawasan adalah salah satu pendekatan dalam *machine learning* di mana sistem dilatih pada data yang tidak memiliki label atau anotasi. Tidak seperti *supervised learning* yang memerlukan data latih berlabel untuk menghasilkan prediksi, *unsupervised learning* bekerja dengan menganalisis struktur atau pola tersembunyi dalam data secara mandiri. Tujuan utama dari metode ini adalah menemukan pola, kelompok (*cluster*), atau representasi yang bermakna tanpa campur tangan manusia dalam proses pelabelan data (Sy et al., 2024).

Unsupervised learning diterapkan pada sistem keamanan berbasis pengenalan wajah dengan kamera IP. Sistem menggunakan metode *image clustering* untuk membedakan antara wajah yang dikenali dan tidak dikenali, tanpa memerlukan label awal. Model ini memungkinkan klasifikasi otomatis berdasarkan kemiripan ciri visual, sehingga cocok untuk sistem keamanan yang memerlukan efisiensi dalam pengelompokan data wajah yang terus-menerus berubah (Desta Yolanda et al., 2021).

Pendekatan ini menunjukkan bagaimana *unsupervised learning* dapat digunakan untuk tugas-tugas praktis dalam lingkungan dunia nyata, seperti pemantauan keamanan otomatis dan penghematan media penyimpanan.

6. *Pseudo-Labeling*

Pseudo-labeling adalah teknik dalam pembelajaran *semi-supervised* yang bertujuan untuk meningkatkan kinerja model dengan memanfaatkan data yang tidak berlabel. Teknik ini dilakukan dengan cara melatih model awal menggunakan data berlabel, lalu model tersebut digunakan untuk memprediksi label pada data yang tidak berlabel. Hasil prediksi ini, yang disebut *pseudo-label*, kemudian digabungkan dengan data asli yang berlabel untuk membentuk dataset baru yang lebih besar. Dataset gabungan ini digunakan kembali untuk melatih ulang model agar lebih akurat dan mampu memanfaatkan data yang awalnya tidak dimanfaatkan secara penuh (Afriyani et al., 2024).

Dalam konteks pemrosesan laporan radiologi, *pseudo-labeling* yang dihasilkan oleh *Large Language Models* (LLMs) terbukti efektif untuk tugas klasifikasi teks medis yang kompleks. Seperti dijelaskan dalam penelitian oleh Wong dan Tanaka. (2025), kerangka kerja DeBERTa-RAD memanfaatkan LLM canggih sebagai guru untuk menghasilkan *pseudo-label* berkualitas tinggi, termasuk status kepastian (*Present*, *Absent*, *Uncertain*), pada korpus laporan yang luas. Proses ini dilanjutkan dengan *knowledge distillation* ke model DeBERTa-Base, yang meningkatkan akurasi melalui kemampuan arsitektur DeBERTa dalam menangkap ketergantungan posisi relatif dan nuansa bahasa klinis seperti negasi dan ketidakpastian. Pendekatan dua tahap ini tidak hanya mengatasi keterbatasan data berlabel manual tetapi juga mencapai skor F1 makro tertinggi (0.912) pada dataset MIMIC-500, mengungguli metode berbasis aturan (*rule-based*) dan model *transformer* lain. Hasil ini menegaskan bahwa *pseudo-labeling* berbasis LLM, dikombinasikan dengan distilasi pengetahuan, menjadi solusi efisien untuk klasifikasi teks medis berskala

besar, khususnya dalam konteks sumber daya berlabel terbatas dan kebutuhan kecepatan inferensi praktis.

7. *Semantic Similarity*

Semantic similarity merupakan pendekatan dalam bidang *Natural Language Processing* (NLP) yang bertujuan untuk mengukur tingkat kemiripan makna antara dua unit teks, seperti kalimat atau dokumen. Tidak seperti pendekatan berbasis pencocokan kata secara langsung, pendekatan ini mampu menangkap konteks makna yang lebih dalam dengan merepresentasikan teks ke dalam bentuk vektor berdimensi tinggi. Salah satu metode yang umum digunakan untuk menghasilkan representasi ini adalah model *Sentence-BERT* (SBERT), yang memanfaatkan arsitektur *transformer* untuk menghasilkan *embedding* kalimat yang bermakna secara semantik (Irbah salsabila dan Yuliant Sibaroni, 2021; Van Son et al., 2021).

Salah satu metode paling umum untuk menghitung kesamaan semantik antara dua vektor representasi teks adalah *cosine similarity*, yang mengukur besar sudut antar dua vektor dalam ruang berdimensi tinggi. Perhitungan *cosine similarity* antara dua vektor A dan B secara matematis didefinisikan pada Rumus 3:

$$\text{cosSim}(A, B) = \frac{A \times B}{\|A\| \times \|B\|} \quad (3)$$

Rumus ini menghasilkan nilai antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa kedua vektor memiliki arah yang sangat mirip atau memiliki makna yang sangat serupa. Pendekatan berbasis SBERT yang dikombinasikan dengan *cosine similarity* telah terbukti efektif dalam berbagai tugas NLP seperti pengelompokan dokumen, pencocokan teks, dan pengukuran kemiripan antar kalimat (Hambi dan Benabbou, 2020).

B. Penelitian Terkait

Peneliti memperoleh banyak inspirasi dan referensi untuk penyusunan skripsi ini dari penelitian sebelumnya, terkait dengan latar belakang masalah pada skripsi ini. Berikut ringkasan penelitian terkait yang menjadi landasan studi ini, disajikan pada Tabel 1:

Tabel 1. Penelitian Terkait

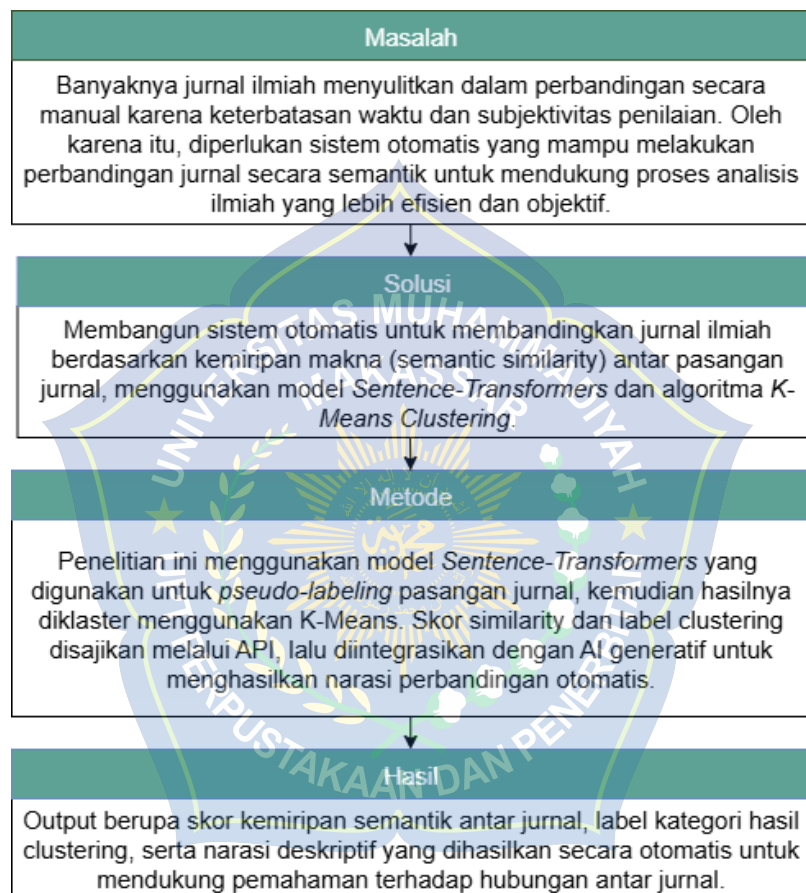
Penelitian	Tujuan/Kasus	Metode	Hasil
<i>Exploring Community Awareness of Mangrove Ecosystem Preservation through Sentence-BERT and K-Means Clustering</i> (Kusumaningrum et al., 2024)	Penelitian ini bertujuan menganalisis opini publik terhadap pelestarian ekosistem mangrove menggunakan data Twitter berbahasa Indonesia. Untuk mengatasi keterbatasan LDA dalam menangkap konteks semantik, peneliti menerapkan Sentence-BERT dan K-Means sebagai pendekatan utama dalam pemodelan topik.	<i>Sentence-BERT, K-Means Clustering</i>	Hasilnya menunjukkan bahwa gabungan SBERT dan K-Means mampu mengidentifikasi sembilan topik utama secara lebih kontekstual dan efisien dibanding LDA, serta memberikan gambaran tren kesadaran masyarakat yang relevan.
<i>A two-phase plagiarism detection system</i>	Studi ini mengembangkan sistem pendeteksian	<i>Multi-layer LSTM, Sentence</i>	Model menghasilkan <i>F1-score</i> sebesar

<i>based on multi-layer lstm networks</i> (Van Son et al., 2021)	plagiarisme dua tahap untuk mengatasi kasus plagiarisme literal dan cerdas. Sistem ini menilai kemiripan semantik antar kalimat dan frasa dengan kombinasi <i>embedding</i> dan <i>deep learning</i> untuk hasil yang lebih presisi.	<i>Embedding</i> (SBERT), <i>Cosine Similarity</i>	94,26% pada dataset PAN 2014. Penggunaan SBERT efektif dalam mengenali plagiarisme kompleks yang sulit dideteksi oleh pendekatan leksikal atau <i>string matching</i> konvensional.
<i>A deep learning based technique for plagiarism detection: A comparative study</i> (Hambi dan Benabbou, 2020)	Penelitian ini membandingkan teknik representasi semantik berbasis <i>Word2Vec</i> dan <i>Doc2Vec</i> dalam konteks deteksi plagiarisme ide, yang dianggap lebih sulit dibanding plagiarisme teks langsung. Fokusnya adalah pada evaluasi kedalaman makna yang bisa ditangkap oleh masing-masing metode.	<i>Word2Vec</i>, <i>Doc2Vec</i>, <i>Cosine Similarity</i>	Hasil studi menunjukkan bahwa pendekatan berbasis <i>word-level</i> belum cukup untuk mendeteksi kemiripan semantik kompleks. Diperlukan model lanjutan seperti SBERT untuk memperoleh hasil yang lebih akurat dan representatif.

<i>Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity</i> (Irbah salsabila dan Yuliant Sibaroni, 2021)	Penelitian ini menganalisis sentimen multi-aspek dari ulasan produk kecantikan untuk mengetahui persepsi pengguna terhadap aspek harga, kemasan, dan aroma. Penambahan <i>semantic similarity</i> digunakan untuk meningkatkan ketepatan klasifikasi opini.	SVM, <i>Semantic Similarity</i> , TF-IDF	<i>Semantic similarity</i> terbukti meningkatkan performa klasifikasi sentimen: 93% untuk harga, 92% kemasan, dan 86% aroma. Metode ini mampu mengatasi variasi kata yang berbeda namun bermakna serupa.
<i>Sentence embedding to improve rumour detection performance model</i> (Anggrainingsih et al., 2024)	Tujuan utama studi ini adalah meningkatkan performa deteksi rumor di Twitter dengan membandingkan model berbasis <i>embedding</i> (BERT) dan pendekatan tradisional berbasis fitur manual.	BERT <i>Sentence Embedding</i> , <i>Neural Network</i>	Model BERT menunjukkan peningkatan akurasi sekitar 10% dibanding pendekatan berbasis <i>feature engineering</i> , serta mengurangi ketergantungan pada proses ekstraksi fitur manual yang rumit dan lambat.

C. Kerangka Berpikir

Kerangka berpikir merupakan model konseptual tentang bagaimana teori berhubungan dengan berbagai faktor yang telah diidentifikasi sebagai masalah penting. Visualisasi kerangka penelitian secara lengkap ditampilkan pada Gambar 1.



Gambar 1. Kerangka Berpikir

BAB III

METODE PENELITIAN

A. Tempat dan Waktu Penelitian

Penelitian ini dilaksanakan di Laboratorium Informatika, Fakultas Teknik, Universitas Muhammadiyah Makassar. Proses pengembangan, pelatihan model, serta pengujian sistem dilakukan menggunakan perangkat komputer pribadi. Pengumpulan dataset dilakukan melalui akses *Application Programming Interface* (API) dari *Semantic Scholar* secara daring.

Adapun waktu pelaksanaan penelitian ini berlangsung selama 2 bulan, dimulai dari bulan Mei 2025 hingga Juli 2025. Rangkaian kegiatan penelitian meliputi tahap perencanaan, persiapan data, implementasi metode *Sentence-Transformers* dan *K-Means Clustering*, pengujian performa sistem, serta analisis hasil untuk mendukung perbandingan semantik antar jurnal ilmiah.

B. Alat dan Bahan

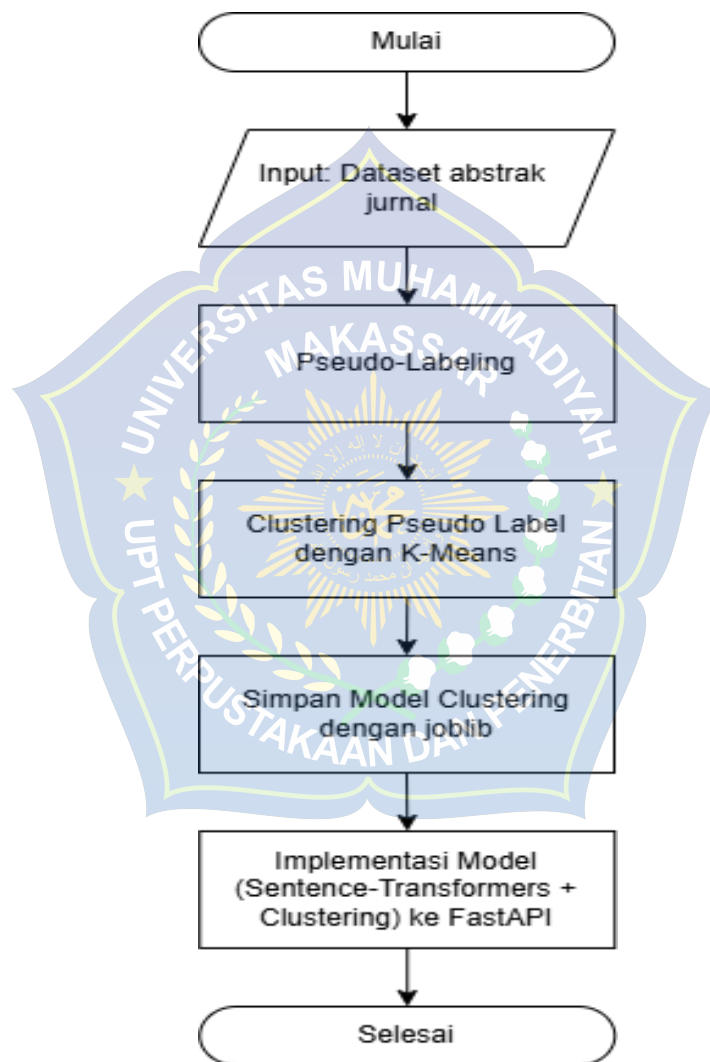
Adapun alat dan bahan yang akan digunakan dalam penelitian ini, yaitu:

1. Kebutuhan *Hardware* (Perangkat Keras)
 - a. Laptop ASUS X409FAC_A409FA
 - b. RAM 12288 MB
2. Kebutuhan *Software* (Perangkat Lunak)
 - a. *Visual Studio Code*
 - b. *Excell*
 - c. *Python*
 - d. *NodeJS*

C. Perancangan Sistem

Untuk mempermudah proses pembuatan dan pengembangan aplikasi, peneliti menggunakan *flowchart*. Dengan demikian, pembuatan aplikasi dapat dilakukan secara terstruktur.

1. Diagram alir (*flowchart*) merupakan representasi visual prosedur sistematis yang memetakan langkah-langkah operasional dan poin keputusan dalam suatu program. Setiap iterasi (*loop*) diwakili oleh struktur simbolis terhubung secara hierarkis. Dalam penelitian ini, Gambar 2 memvisualisasikan alur perancangan model dan klusterisasi, sedangkan Gambar 3 memvisualisasikan proses pengolahan jurnal dan generasi output.



Gambar 2. *Flowchart* Perancangan Model dan *Clustering*

Penjelasan pada Gambar 2 adalah sebagai berikut:

a) Mulai

Proses dimulai dengan inisialisasi sistem untuk melakukan *pseudo-labeling* menggunakan model *Sentence-Transformers*.

b) Input: Dataset Abstrak Jurnal

- Dataset terdiri dari pasangan data berisi kolom: title1, abstract1, title2, abstract2
- Data ini akan menjadi objek pembandingan menggunakan model *Sentence-Transformers*.

c) *Pseudo-Labeling*

- Model *Sentence-Transformers* dimanfaatkan untuk menghitung skor kemiripan semantik antar pasangan abstrak jurnal dalam dataset.
- Hasilnya digunakan untuk memberi label kemiripan secara otomatis, yang disebut sebagai *pseudo label*.

d) *Clustering Pseudo Label* dengan *K-Means*

- *Pseudo label* yang dihasilkan dikelompokkan menggunakan algoritma *Clustering K-Means*.
- *Clustering* digunakan untuk mengidentifikasi pola kemiripan semantik antar jurnal.

e) Simpan Model *Clustering* dengan *Joblib*

Model hasil pelatihan *clustering* disimpan menggunakan *joblib* agar dapat digunakan kembali di sistem produksi (API).

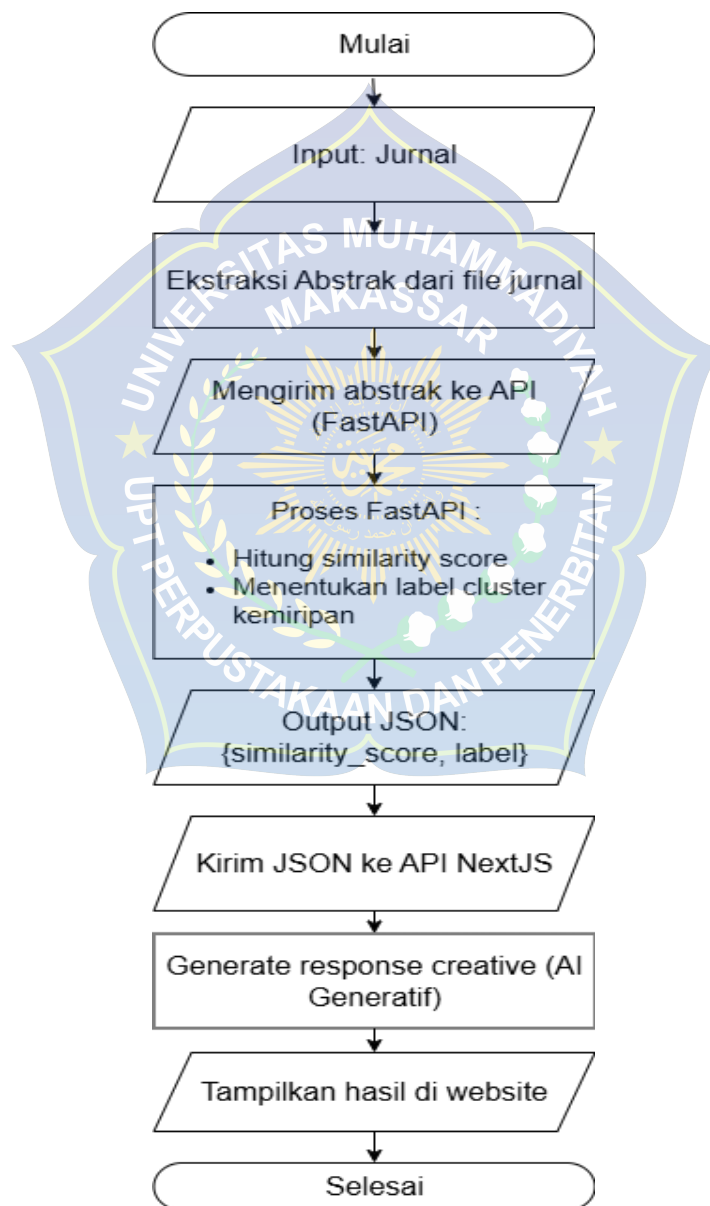
f) Implementasi Model (*Sentence-Transformers* dan *Clustering*) ke FastAPI

- Model *Sentence-Transformers* terbaik dan model *clustering* diintegrasikan ke dalam REST API menggunakan *framework* FastAPI.
- API bertugas menerima input teks dan mengembalikan skor kemiripan serta label klaster.

g) Selesai

Proses perancangan sistem selesai dan siap digunakan untuk integrasi dengan antarmuka pengguna atau sistem lainnya.

Setelah sistem inti (*Sentence-Transformers* dan *K-Means*) terintegrasi dalam API FastAPI, diperlukan mekanisme operasional untuk menerima input pengguna dan menghasilkan output yang interpretatif. Pada Gambar 3 memvisualisasikan alur kerja sistem ketika pengguna mengunggah dokumen jurnal, dimulai dari ekstraksi teks hingga generasi narasi analitis. Proses ini dirancang untuk memastikan interaksi yang efisien antara komponen teknis dan kebutuhan praktis peneliti.



Gambar 3. *Flowchart* Pengolahan Jurnal dan Output

Penjelasan Gambar 3 adalah sebagai berikut:

a) Mulai

Titik awal ketika pengguna ingin membandingkan dua jurnal ilmiah.

b) Input *File* atau Dokumen Jurnal

Pengguna mengunggah *file* jurnal, dalam format PDF.

c) Ekstraksi Abstrak dari *File* Jurnal

Sistem membaca isi *file* dan mengekstrak bagian abstrak dari kedua jurnal sebagai input utama untuk analisis semantik.

d) Mengirim Abstrak ke API (FastAPI)

Abstrak yang sudah diekstrak dikirimkan ke backend API (dibuat dengan FastAPI) untuk diproses lebih lanjut.

e) Proses FastAPI

Server menjalankan dua proses utama:

- Hitung *Similarity Score*: Menggunakan model terbaik dari *Sentence-Transformers*, API menghitung skor kemiripan antar abstrak.
- Menentukan Label *Cluster* Kemiripan: Berdasarkan model *K-Means* hasil *pseudo-labeling*, sistem mengklasifikasikan pasangan jurnal ke dalam label kluster misalnya Rendah, Sedang, dan Tinggi.

f) Output JSON

Pada tahap ini, setelah melewati bagian proses pada FastAPI maka akan mengembalikan Output berupa data JSON, misalnya:

```
{
  "similarity_score": (Rentang score dari -1 hingga 1),
  "label_kemiripan": (Rendah, Sedang, Tinggi)
}
```

g) Kirim JSON ke API Next.js

Data hasil JSON diteruskan dari backend FastAPI ke frontend yang menggunakan framework Next.js.

h) *Generate Response Creative* (AI Generatif)

Berdasarkan skor kemiripan dan label kluster, sistem menghasilkan

narasi perbandingan kreatif menggunakan AI.

i) Tampilkan Hasil di Website

Narasi hasil analisis ditampilkan pada antarmuka website, lengkap dengan informasi skor dan interpretasi klaster.

j) Selesai

Proses selesai, pengguna dapat membaca dan menyimpan hasil perbandingan.

D. Teknik Pengujian Sistem

Pada penelitian ini, teknik pengujian yang dilakukan menggunakan pendekatan *Black-Box Testing*. Metode ini berfokus pada verifikasi fungsionalitas eksternal sistem tanpa menganalisis struktur internal kode, dengan prinsip utama mengamati hubungan antara *input* dan *output* (Ayuningtyas et al., 2023). Sesuai karakteristiknya, *black-box* testing ideal untuk validasi:

1. Kesesuaian sistem dengan kebutuhan pengguna (*user requirements*)
2. Kelayakan alur bisnis (*business process flow*)
3. Respons sistem di bawah beban kerja tertentu (*performance under load*)

Dalam konteks penelitian ini, metode ini diterapkan untuk memastikan fungsi inti sistem yaitu perhitungan skor kemiripan semantik dan klasifikasi klaster beroperasi sesuai ekspektasi desain.

E. Teknik Analisis Data

Teknik analisis data adalah proses mempelajari dan mengolah data untuk mengidentifikasi pola, hubungan, dan informasi penting yang terkandung di dalamnya. Tujuannya adalah untuk mendapatkan pemahaman yang lebih dalam tentang data yang dianalisis dan mengambil keputusan berdasarkan informasi yang ditemukan. Pemilihan teknik analisis data yang tepat juga tergantung pada jenis data yang dikumpulkan.

Untuk mencapai hasil yang dilakukan peneliti melakukan beberapa tahapan pengolahan data sebelum dilakukan perhitungan dan analisa dengan metode

yang ditentukan. Adapun Langkah-langkah analisis data yaitu sebagai berikut:

1. Pra-pemrosesan Data

Pra-pemrosesan data merupakan tahap awal yang sangat penting dalam analisis data, karena di sinilah data melalui pembersihan, transformasi, dan normalisasi, data yang semula beragam dan tidak konsisten dapat diubah menjadi format yang seragam, yang mendukung konvergensi model *machine learning* serta meningkatkan keakuratan hasil analisis.

2. Ekstraksi Fitur

Langkah selanjutnya ialah ekstraksi fitur, ekstraksi fitur merupakan proses mengubah data teks menjadi representasi numerik yang dapat diproses oleh model *machine learning*, dalam penelitian ini adalah *embedding* dari *Sentence-Transformers*.

3. *Pseudo-Labeling*

Konsep *Pseudo-labeling* adalah teknik yang digunakan dalam *machine learning*, terutama dalam konteks *supervised* dan *semi-supervised learning*, di mana model pembelajaran mesin digunakan untuk memberikan label prediksi pada data tanpa label, dan label-label ini disebut *pseudo-labels*.

4. Analisis Klaster

Setelah data melalui proses *self-training* dengan *pseudo labeling*, langkah selanjutnya adalah analisis klaster. Teknik ini digunakan untuk mengelompokkan data berdasarkan kesamaan fitur sehingga struktur atau pola tersembunyi dalam data dapat terungkap.

5. Evaluasi dan Interpretasi Hasil

Tahap akhir dari proses analisis adalah evaluasi dan interpretasi hasil, yang merupakan langkah krusial untuk memastikan bahwa output analisis merefleksikan struktur sebenarnya dalam data. Penggunaan metrik seperti *Silhouette Score* memberikan ukuran numerik atas kualitas hasil analisis. Hasil metrik ini selanjutnya diinterpretasikan untuk menentukan tingkat keandalan klaster dan kesesuaian model sebelum disajikan melalui AI generatif.

BAB IV

HASIL DAN PEMBAHASAN

A. Dataset Jurnal

Dataset dalam penelitian ini dikumpulkan melalui *Semantic Scholar* API dengan tiga topik, yakni teknologi, kesehatan, dan ekonomi. Proses kurasi data menerapkan kriteria ketat dimana setiap judul jurnal harus mengandung minimal 4 kata untuk menjamin deskriptivitas memadai, sementara abstrak wajib terdiri atas setidaknya 50 kata guna memastikan kedalaman analitis yang memenuhi standar akademik.

Sebanyak 200 abstrak jurnal berhasil terkumpul dengan periode publikasi 2020-2025. Selanjutnya, 200 abstrak dipasangkan melalui proses pengacakan sistematis menghasilkan 100 pasangan komparatif yang siap untuk analisis lebih lanjut. Gambar 4 menampilkan sampel 19 baris pertama struktur dataset, mengilustrasikan format empat kolom utama yang mencakup pasangan judul dan abstrak jurnal: *title1*, *abstract1*, *title2*, dan *abstract2*.

	A	B	C	D
1	title1	abstract1	title2	abstract2
2	The Health Utilities Index (HUI): conc	This is a review of the Health Utilities	Perceived Usefulness, Ease of Use, an	This paper presents the findings of tv
3	Internet paradox. A social technology	The Internet could change the lives of	Ecological momentary interventions: PURPOSE	Psychosocial and health bel
4	Acculturation and Latino health in th	This chapter provides an overview of	Dissecting racial bias in an algorithm (	Racial bias in health algorithms The U
5	Best Practices for Developing and Val	Scale development and validation are 'Old	Economy' Inputs for 'New Econo	This paper discusses the results of a t
6	Contamination of drinking-water by ;	The contamination of groundwater b	Linear Economy Versus Circular Econ	Abstract Upon visiting the existing lite
7	Collaborative Economy and Tourism: House	swapping, ridesharing, volunte	Technology development for the proc	A biorefinery that supplements its m
8	Toward a Greater Understanding of !	The informal economy consists of bu	Evaluating the public health impact of	Progress in public health and commu
9	Shaping Agility through Digital Optio	Agility is vital to the innovation and c	Police Brutality and Black Health: Sett	We investigated links between police
10	Environment, economy and society: f	Sustainable development is a contest	Electronic Data Interchange and Smal	Many EDI researchers and practition
11	Can electronic medical record system	To broadly examine the potential hea	The world economy : historical statist	Angus Maddison provides a compreh
12	The Relationship between Social Coh	Social cohesion involves the interspers	Technology as an occasion for structu	New medical imaging devices, such as
13	The health benefits of urban green s	BACKGROUND Urban development pi	The Duality of Technology: Rethinking	This paper develops a new theoretica
14	Perceived Usefulness, Perceived Ease	Valid measurement scales for predict	Chains and networks, territories and	:A vast and continually expanding liter
15	What Do We Learn from the Weathe	A rapidly growing body of research a	Hispanic health in the USA: a scoping	Hispanics are the largest minority gro
16	What the evidence shows about pati	Patient engagement is an increasingly	Mental Health Smartphone Apps: Rev	Background The number of mental h
17	Trasformismo or transformation? Th	Abstract What does IPE have to contr	Photovoice: a participatory action res	Photovoice is a participatory action re
18	The Impact of Knowledge-Based Eco	The aim of this paper is to examine th	Capturing the Complexity in Advance	The past decade has brought advanc
19	Socioeconomic disparities in health i	OBJECTIVES We aimed to describe so	Does health literacy mediate the relat	While socioeconomic disparities are
20	Rethinking the Informal Economy: Lir	This paper explores the relationship	Antecedents of trust in the sharing ec	Users and potential users of the shar

Gambar 4. Sampel Dataset Jurnal (19 Baris Pertama)

B. Pseudo-Labeling Dataset Jurnal

Pada tahap ini, akan memakai model *all-mpnet-base-v2* dari *Sentence-Transformers*, Keunggulan utama model ini terletak pada kemampuannya dalam merepresentasikan teks panjang seperti paragraf atau abstrak secara lebih kontekstual, Selanjutnya proses *pseudo-labeling* menghasilkan distribusi skor *cosine similarity* yang dianalisis melalui visualisasi histogram dengan *Kernel Density Estimation* (KDE).

1. Proses Pseudo-Labeling

- Ekstraksi Abstrak Unik: Abstrak dari kolom *abstract1* dan *abstract2* diidentifikasi secara unik untuk menghindari redundansi komputasi.
- Pembentukan *Embedding*: Setiap abstrak diubah menjadi vektor berdimensi tinggi menggunakan model *Sentence-Transformers*.
- Perhitungan *Cosine Similarity*:

Pada dataset baris pertama vektor *embedding* setelah normalisasi L2 diperoleh:

$$\|v_1\| = 1.0000, \quad \|v_2\| = 1.0000$$

Produk titik (*dot product*) kedua vektor terhitung:

$$v_1 \times v_2 = 0.1802398$$

Maka, *cosine similarity* (*pseudo-label*) baris pertama adalah:

$$\text{cosSim} = \frac{v_1 \times v_2}{\|v_1\| \times \|v_2\|} = \frac{0.1802398}{1.0000 \times 1.0000} = 0.1802398$$

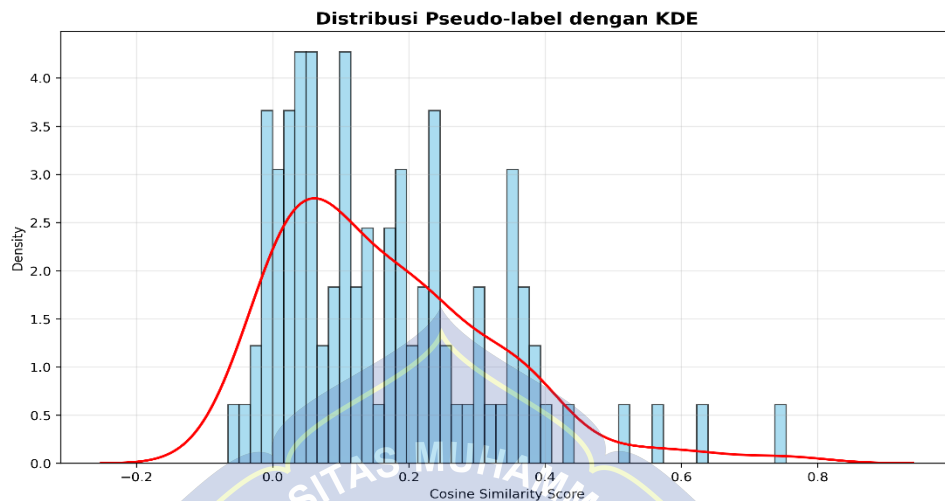
Sehingga, nilai *pseudo-label* untuk baris pertama pada dataset adalah 0.1802398.

2. Distribusi Skor Pseudo-Label

Visualisasi Histogram KDE pada Gambar 5 mengungkap karakteristik kunci distribusi skor:

- Bentuk Distribusi: Multimodal dengan puncak dominan di rentang 0.0–0.2 dan minor di 0.5–0.6, menunjukkan variasi tingkat kemiripan yang signifikan.

- Konsentrasi Data: Sebagian besar pasangan jurnal (67%) terkonsentrasi pada rentang skor 0.0–0.25, sementara sedikit pasangan (4%) berada di rentang 0.5–1.00.



Gambar 5. Distribusi Skor *Pseudo-Label* (Histogram & KDE)

3. Tampilan Dataset

Sebelum tahap *pseudo-labeling* dapat dilihat pada Gambar 6

	A	B	C	D
1	title1	abstract1	title2	abstract2
2	The Health Utilities Index (HUI): conc This is a review of the Health Utilities Perceived Usefulness, Ease of Use, an This paper presents the findings of tw			
3	Internet paradox. A social technology The Internet could change the lives o Ecological momentary interventions: PURPOSE Psychosocial and health bel			
4	Acculturation and Latino health in the This chapter provides an overview of Dissecting racial bias in an algorithm Racial bias in health algorithms The U			
5	Best Practices for Developing and Val Scale development and validation are 'Old Economy' Inputs for 'New Econo This paper discusses the results of a t			
6	Contamination of drinking-water by e The contamination of groundwater b Linear Economy Versus Circular Econ Abstract Upon visiting the existing lite			

Gambar 6. Dataset sebelum proses *Pseudo-Labeling*

Sesudah melalui tahap *pseudo-labeling* ditampilkan pada Gambar 7

	A	B	C	D	E
1	title1	abstract1	title2	abstract2	pseudo-label
2	The Health Utilities Index (HUI): conc This is a review of the Health Utilities Perceived Usefulness, Ease of Use, ar This paper presents the findings of tw				0.180239886
3	Internet paradox. A social technology The Internet could change the lives o Ecological momentary interventions: PURPOSE Psychosocial and health bel				0.391628027
4	Acculturation and Latino health in the This chapter provides an overview of Dissecting racial bias in an algorithm Racial bias in health algorithms The U				0.371246964
5	Best Practices for Developing and Val Scale development and validation are 'Old Economy' Inputs for 'New Econo This paper discusses the results of a i				0.054675169
6	Contamination of drinking-water by e The contamination of groundwater b Linear Economy Versus Circular Econ Abstract Upon visiting the existing lite				0.10084822

Gambar 7. Dataset sesudah proses *Pseudo-Labeling*

C. Evaluasi Kluster Optimal

Pada tahap ini dilakukan evaluasi terhadap hasil *pseudo-labeling* yang telah diterapkan pada dataset jurnal, guna menentukan jumlah kluster (k) yang paling representatif. Evaluasi ini bertujuan untuk mengidentifikasi nilai k yang mampu menghasilkan pengelompokan data yang efisien serta memiliki validitas internal yang baik.

1. Metodologi Evaluasi

Penentuan jumlah kluster optimal dilakukan dengan tiga metrik evaluasi internal, yaitu *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index* (DBI). Ketiganya digunakan untuk menilai kualitas segmentasi hasil klusterisasi terhadap data *pseudo-label*.

1) *Elbow Method*

Metode ini digunakan untuk mengevaluasi nilai *inertia* atau *within-cluster sum of squares* (WCSS) dari berbagai nilai k . Nilai *inertia* yang lebih rendah menunjukkan bahwa objek-objek dalam suatu kluster memiliki kedekatan yang lebih tinggi, sehingga semakin baik. Titik siku atau *elbow* pada grafik merupakan indikator jumlah kluster optimal, di mana penurunan *inertia* mulai melambat secara signifikan.

2) *Silhouette Score*

Metode ini mengukur seberapa baik setiap objek cocok dengan klasternya dibandingkan dengan kluster lain yang paling dekat. Nilai *Silhouette Score* berkisar dari -1 hingga $+1$, dengan nilai mendekati $+1$ menunjukkan pembentukan kluster yang baik.

3) *Davies-Bouldin Index* (DBI)

DBI menilai rasio antara sebaran dalam kluster dengan jarak antar kluster. Nilai yang lebih kecil menunjukkan kluster yang lebih terpisah dan lebih homogen. DBI memberikan perspektif kuantitatif tambahan dalam menilai efektivitas struktur kluster.

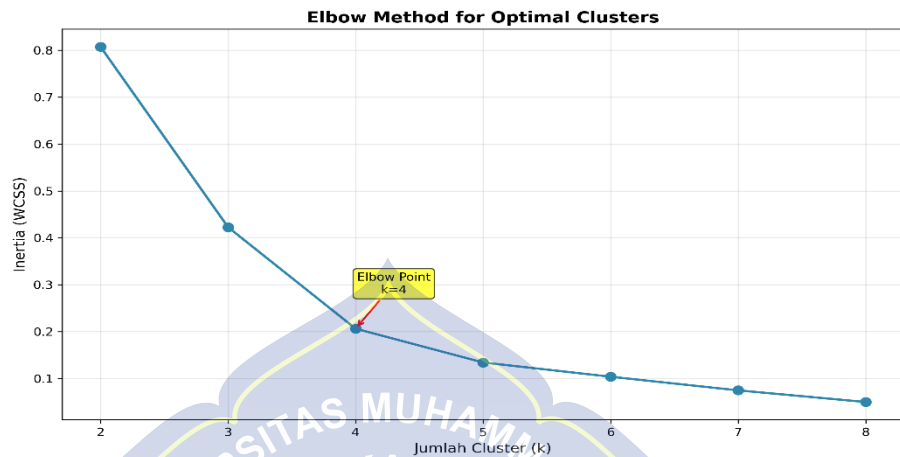
Ketiga metrik ini dihitung untuk rentang nilai k dari 2 hingga 8 menggunakan *K-Means*. Hasilnya menjadi dasar untuk menentukan jumlah kluster yang paling sesuai secara statistik dan semantik.

2. Hasil Evaluasi dan Interpretasi

a) *Elbow Method*

Gambar 8 memperlihatkan tren penurunan nilai *inertia* seiring bertambahnya jumlah kluster. Penurunan yang cukup tajam terlihat dari $k = 2$ hingga $k = 4$, sedangkan setelah $k = 4$ laju penurunannya menjadi

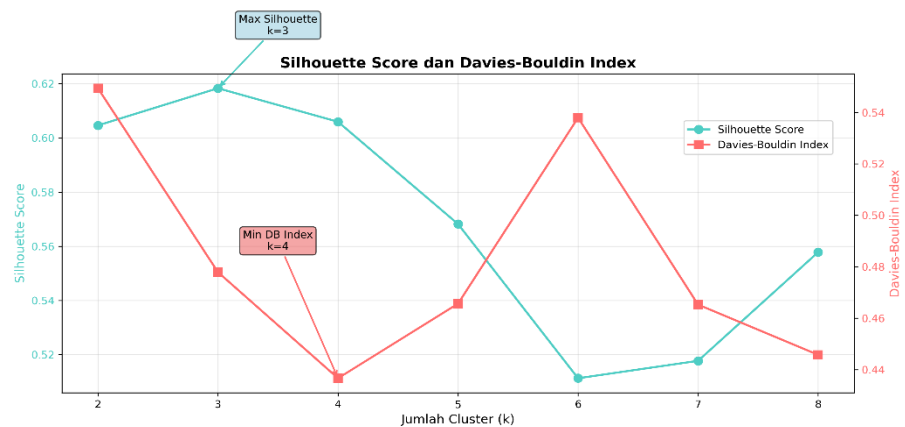
lebih landai. Pola ini membentuk titik siku (*elbow point*) pada $k = 4$, yang menunjukkan bahwa penambahan jumlah kluster setelah titik tersebut tidak memberikan peningkatan signifikan dalam homogenitas data.



Gambar 8. Grafik *Elbow Method* dari *Pseudo-Label*

b) *Silhouette Score* dan *Davies–Bouldin Index* (DBI)

Gambar 9 menampilkan dua kurva evaluasi, yaitu *Silhouette Score* dan DBI, untuk $k = 2$ hingga $k = 8$. Nilai *Silhouette Score* tertinggi diperoleh pada $k = 3$, sedangkan DBI terendah terdapat pada $k = 4$. Hal ini menunjukkan bahwa $k = 3$ memiliki pemisahan kluster terbaik berdasarkan *Silhouette Score*, sementara $k = 4$ memiliki sebaran internal yang paling kompak dan jarak antar kluster yang relatif besar menurut DBI.



Gambar 9. Grafik *Silhouette Score* (Biru) dan DBI (Jingga) dari *Pseudo-Label*

Untuk memberikan gambaran kuantitatif secara menyeluruh, hasil perhitungan ketiga metrik evaluasi terhadap setiap variasi jumlah klaster disajikan pada Tabel 2 berikut:

Tabel 2. Nilai Metrik Evaluasi untuk Setiap Jumlah Klaster

Jumlah <i>Cluster</i> (k)	<i>Inertia</i> (WCSS)	<i>Silhouette</i> <i>Score</i>	<i>Davies-</i> <i>Bouldin Score</i>
2	0.81	0.6046	0.5494
3	0.42	0.6183	0.4779
4	0.21	0.6060	0.4366
5	0.13	0.5682	0.4656
6	0.10	0.5113	0.5380
7	0.07	0.5177	0.4653
8	0.05	0.5578	0.4457

3. Perbandingan Klaster Optimal

Untuk memperkuat dasar penentuan jumlah klaster optimal, dilakukan perbandingan hasil evaluasi dari ketiga metrik yang telah digunakan, yaitu *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index* (DBI).

Tabel 3 berikut merangkum hasil ketiga metrik terhadap variasi jumlah klaster dari 2 hingga 8:

Tabel 3. Perbandingan Evaluasi Jumlah Klaster

<i>Klaster</i> (k)	<i>Elbow Method</i>	<i>Silhouette Score</i>	<i>Davies-Bouldin</i> <i>Index</i> (DBI)
2	Penurunan <i>inertia</i> signifikan, namun terlalu umum	Skor cukup tinggi, namun pembentukan klaster terlalu kasar.	DBI masih tinggi, menandakan pemisahan antar klaster belum baik.
3	Penurunan <i>inertia</i> masih berlanjut	Skor tertinggi, namun stabilitas antar klaster menurun.	DBI cukup rendah, tapi tidak paling optimal.

4	Penurunan <i>inertia</i> mulai melandai	Skor <i>Silhouette</i> tetap tinggi dan cenderung stabil.	DBI terendah, menunjukkan keseimbangan terbaik antar klaster.
5	Penurunan <i>inertia</i> kurang signifikan	Skor menurun, menunjukkan kualitas klaster mulai menurun.	DBI mulai meningkat kembali.
6	Penurunan <i>inertia</i> sangat kecil	Skor <i>Silhouette</i> lebih rendah, klaster terlalu banyak.	DBI naik drastis, menunjukkan tumpang tindih antar klaster.
7	Penurunan <i>inertia</i> sangat kecil	Skor sedikit membaik, tapi tidak signifikan.	DBI tetap tinggi, pemisahan tidak membaik.
8	Penurunan hampir datar	Skor meningkat sedikit, namun <i>over-clustering</i> .	DBI lebih tinggi dari $k = 4$.

Berdasarkan Tabel 3 di atas, nilai $k = 4$ dipilih sebagai jumlah klaster optimal karena menunjukkan keseimbangan yang baik antara ketiga metrik evaluasi. Pada nilai ini, *inertia* mulai melandai (indikator dari titik siku pada *Elbow Method*), nilai *Silhouette Score* tetap tinggi yang menandakan pembentukan klaster masih stabil, serta nilai DBI yang paling rendah, yang berarti klaster memiliki kekompakan internal dan keterpisahan antar klaster yang sangat baik. Jumlah empat klaster ini juga memberikan fleksibilitas semantik yang cukup untuk mengelompokkan jurnal ke dalam kategori kemiripan yang jelas dan mudah diinterpretasikan.

D. Proses Klastering *Pseudo-Label*

Pada tahap ini bertujuan mengelompokkan data hasil *pseudo-labeling* ke dalam sejumlah klaster dengan tingkat kemiripan semantik antar jurnal. Pengelompokan dilakukan menggunakan algoritma *K-Means Clustering* berdasarkan jumlah klaster optimal dari evaluasi sebelumnya memakai metrik *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index* (DBI).

1. Proses Klasterisasi

Dengan jumlah klaster yang telah ditentukan, data *pseudo-label* selanjutnya diklasterkan menggunakan algoritma *K-Means*. Proses klasterisasi ini menghasilkan empat *centroid* yang diurutkan berdasarkan nilai *pseudo-label*. Selanjutnya, tiap klaster diberi label semantik secara manual dari yang paling rendah hingga paling tinggi tingkat kemiripannya. Adapun label kategorikal yang merefleksikan derajat kemiripan, sebagai berikut:

- Tidak Relevan
- Sedikit Berkaitan
- Cukup Berkaitan
- Sangat Berkaitan

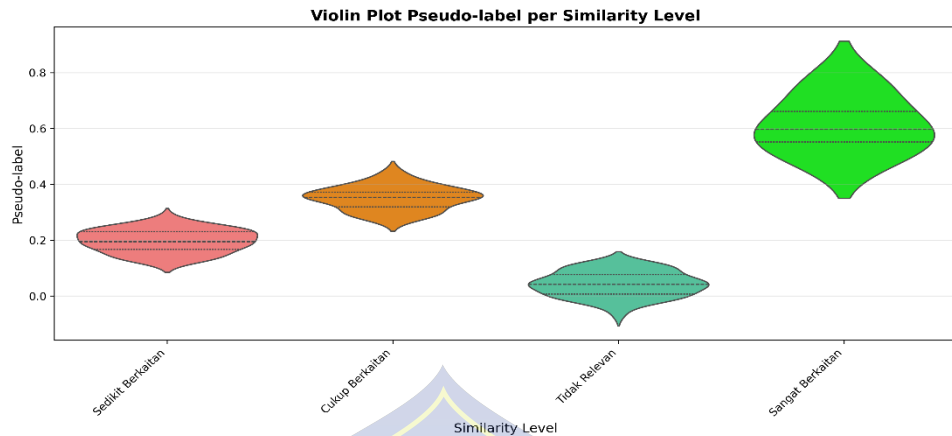
Pemberian label ini bertujuan agar setiap kelompok dapat diinterpretasikan secara semantik dan dimanfaatkan dalam analisis maupun proses naratif yang mendalam pada tahapan sistem lanjutan. Untuk mendukung transparansi metode, Dibawah ini akan memperlihatkan potongan kode yang menunjukkan proses pelatihan model *K-Means* serta pemetaan hasil klaster ke dalam label semantik:

```
kmeans = KMeans(n_clusters=4, random_state=42)
df['cluster'] = kmeans.fit_predict(df[['pseudo-label']])
centers = kmeans.cluster_centers_.flatten()
sorted_idx = np.argsort(centers)
labels = ["Tidak Relevan", "Sedikit Berkaitan", "Cukup Berkaitan", "Sangat Berkaitan"]
cluster_map = {
    int(sorted_idx[i]): labels[i]
    for i in range(len(labels))
}
df['similarity_level'] = df['cluster'].map(cluster_map)
joblib.dump(kmeans, './model/kmeans_model.joblib')
```

2. Visualisasi Distribusi *Pseudo-Label*

Untuk menggambarkan sebaran nilai *pseudo-label* dalam masing-masing klaster, digunakan *violin plot* sebagai visualisasi distribusi data. Visualisasi ini menampilkan kepadatan nilai serta posisi kuartil dalam tiap label kemiripan.

Gambar 10 berikut menunjukkan hasil visualisasi distribusi *pseudo-label* terhadap empat kategori kemiripan:



Gambar 10. Visualisasi klastering *pseudo-label* pada masing-masing kategori

Berdasarkan Visualisasi *Violin Plot* pada Gambar 10, distribusi setiap kategori menunjukkan ciri khas yang konsisten dengan tingkat kemiripannya. Untuk memvalidasi akurasi statistik deskriptif yang dihasilkan oleh program, dilakukan perhitungan manual menggunakan *Microsoft Excel* pada dataset yang sama. Tabel 4 merangkum dan membandingkan statistik deskriptif dari distribusi nilai *pseudo-label* pada masing-masing kategori, baik yang dihasilkan oleh model program maupun perhitungan manual.

Tabel 4. Statistik Deskriptif *Pseudo-Label* per Klaster

Similarity Level	Jumlah Pasangan Jurnal	Sumber	Min	Max	Mean	Median	Std
Tidak Relevan	49	Model	-0.066	0.118	0.042	0.043	0.045
		Excel	-0.065	0.117	0.042	0.042	0.045
Sedikit Berkaitan	29	Model	0.126	0.273	0.197	0.196	0.041
		Excel	0.126	0.272	0.197	0.195	0.040
Cukup Berkaitan	18	Model	0.279	0.435	0.351	0.354	0.042
		Excel	0.279	0.435	0.350	0.353	0.041
Sangat Berkaitan	4	Model	0.509	0.754	0.615	0.598	0.105
		Excel	0.509	0.753	0.614	0.597	0.105

Distribusi *pseudo-label* pada tiap klaster menunjukkan pola konsisten pada tingkat kemiripan semantik antar jurnal. Setiap kategori memiliki rentang nilai yang relatif terpisah dan rata-rata skor yang meningkat dari kategori Tidak Relevan hingga Sangat Berkaitan. Perbandingan pada Tabel 4 menunjukkan bahwa nilai statistik yang dihasilkan oleh model program memiliki konsistensi yang sangat tinggi dengan hasil perhitungan manual di *Excel*. Hal ini memvalidasi bahwa proses komputasi yang dilakukan oleh sistem sudah akurat dan dapat diandalkan.

Keberhasilan validasi ini mencerminkan keberhasilan *pseudo-labeling* dan klasterisasi dalam mengelompokkan pasangan jurnal secara logis dan terstruktur. Selanjutnya, Model *K-Means* disimpan dalam format JOBLIB, sedangkan pemetaan indeks klaster ke label kategorikal disimpan dalam bentuk JSON. Langkah ini memastikan model dan label siap digunakan pada tahap implementasi sistem.

E. Implementasi Sistem

Tahap selanjutnya dalam penelitian ini adalah implementasi sistem berbasis web yang mengintegrasikan model pembelajaran mesin dan antarmuka pengguna (*user interface*) untuk melakukan perbandingan jurnal secara otomatis. Sistem ini dibangun menggunakan arsitektur dua lapis, yaitu *backend Python* berbasis *FastAPI* dan *frontend-backend Next.js* yang terintegrasi dengan layanan AI untuk pembuatan ringkasan.

1. Arsitektur Sistem

Setelah dilakukan proses *pseudo-labeling* dan klasterisasi dengan nilai klaster optimal sebanyak 4 klaster, hasil model digunakan dalam pengembangan *backend* untuk melakukan inferensi terhadap skor kemiripan antar jurnal. *Backend* ini menjalankan pipeline sebagai berikut:

- Ekstraksi teks abstrak dari kedua jurnal yang diunggah pengguna.
- Prapemrosesan teks menggunakan *regex* dan normalisasi.
- Ekstraksi *embedding* semantik menggunakan model *all-mpnet-base-v2*.

- Perhitungan *Cosine Similarity* antara dua *embedding*.
- Prediksi kluster menggunakan model *K-Means*.
- Mapping label semantik berdasarkan hasil klasterisasi.
- Pengembalian hasil skor kemiripan dan label dalam dua bahasa (Indonesia dan Inggris).

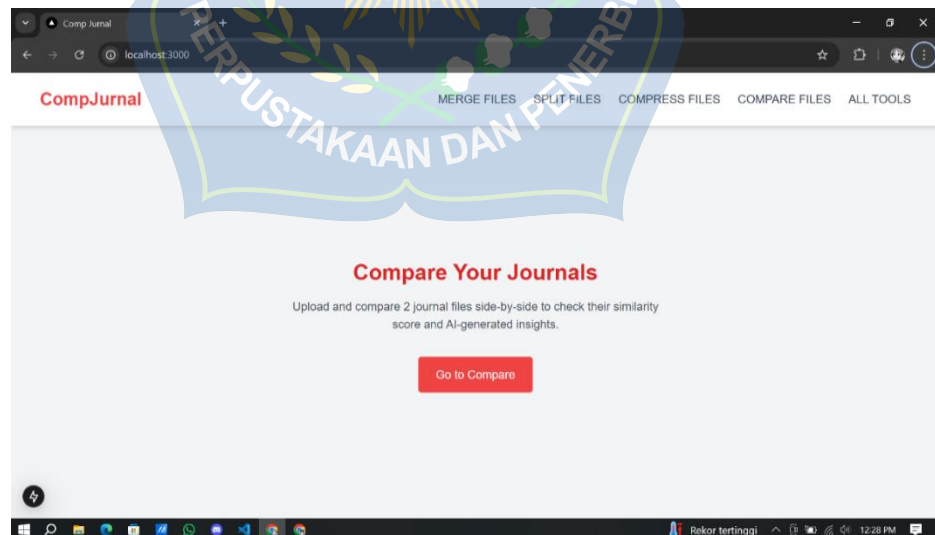
2. Integrasi Antarmuka dan Proses Alur

Antarmuka sistem dibangun dengan *Next.js* yang tidak hanya berfungsi sebagai frontend, tetapi juga sebagai penghubung ke *backend Python* serta penyedia ringkasan berbasis model AI generatif (*gemini*). Ringkasan tersebut dihasilkan berdasarkan konteks skor kemiripan dan label yang diperoleh dari proses *backend Python*.

Adapun alur sistem dapat dijelaskan melalui tahapan implementasi sebagai berikut:

- Halaman Awal Sistem (*Home Page*)

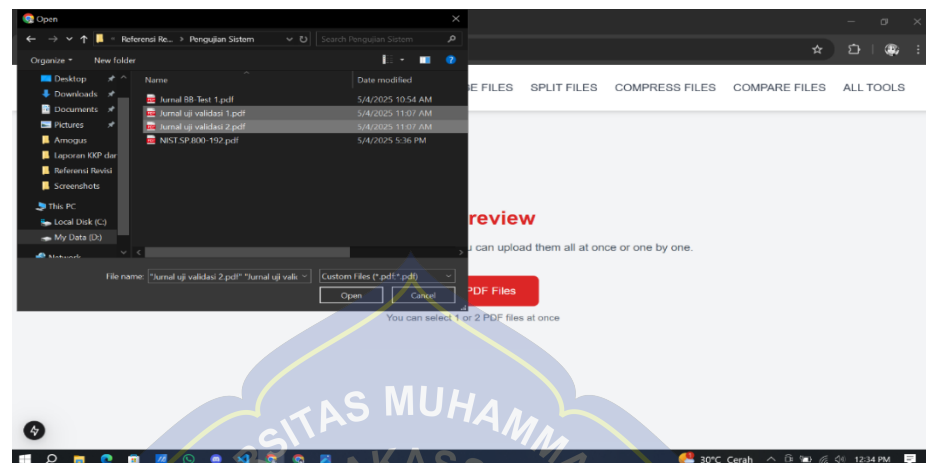
Gambar 11 menunjukkan tampilan awal antarmuka sistem ketika pengguna mengakses halaman utama. Tombol “*Go to Compare*” akan mengarahkan pengguna untuk memulai proses perbandingan jurnal.



Gambar 11. Tampilan halaman utama sistem (*home page*)

- Unggah File Jurnal

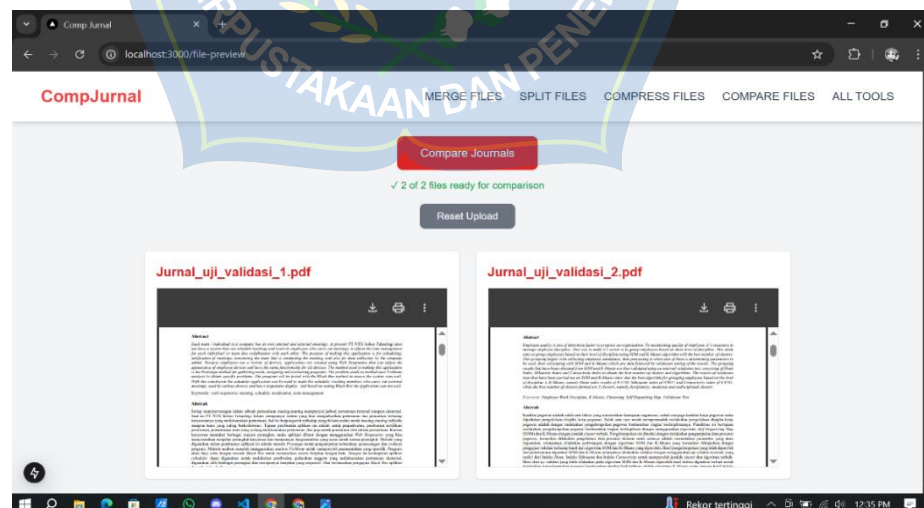
Pengguna diminta untuk mengunggah dua file jurnal dalam format PDF. Sistem hanya mengizinkan tepat dua file untuk dianalisis. Tampilan ini ditunjukkan pada Gambar 12.



Gambar 12. Proses unggah dua jurnal oleh pengguna

- Pratinjau File Jurnal

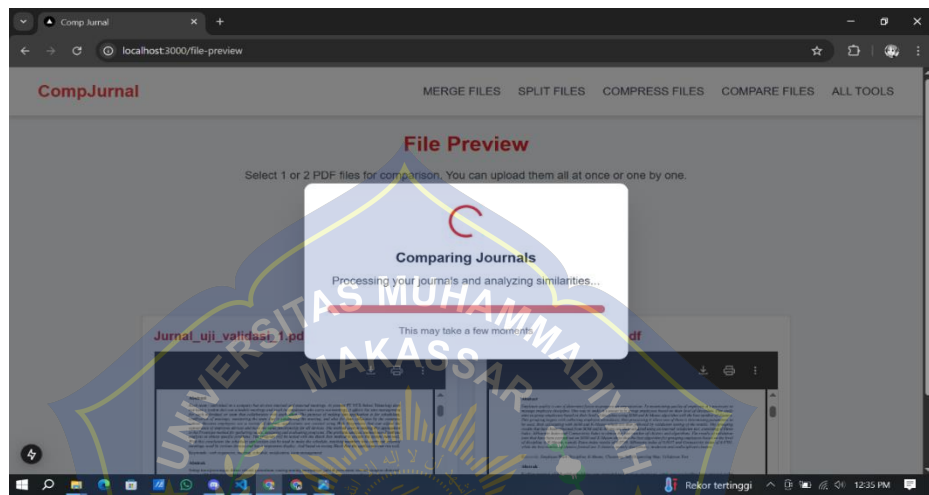
Setelah file berhasil diunggah, sistem akan menampilkan pratinjau kedua jurnal untuk memastikan isi dokumen sebelum dilakukan analisis. Ilustrasi tampilan ini dapat dilihat pada Gambar 13.



Gambar 13. Pratinjau file jurnal sebelum proses komparasi

- Proses Komparasi

Setelah pengguna menekan tombol “*Compare*”, sistem akan mengekstrak abstrak dari kedua jurnal, menghitung skor kemiripan menggunakan *backend Python*, dan mengklasifikasikan hasilnya ke dalam label kategorikal. Gambar 14 menunjukkan proses loading saat sistem menjalankan komputasi.



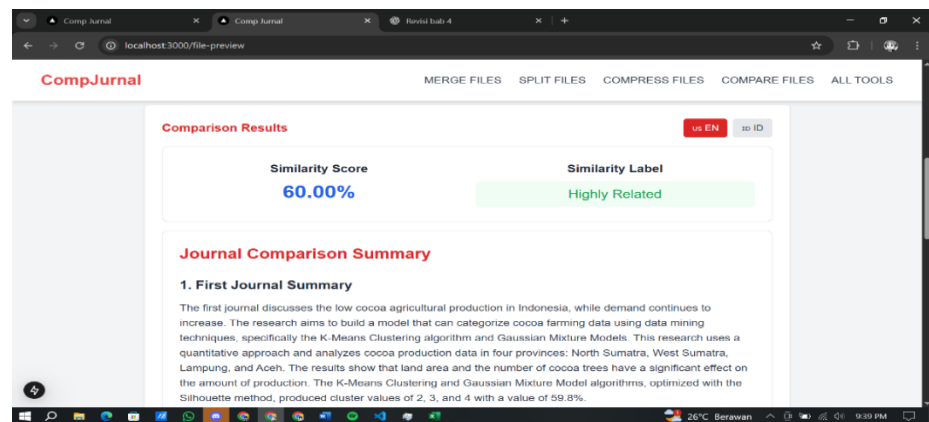
Gambar 14. Proses komputasi sistem dalam melakukan perbandingan

- Tampilan Hasil Komparasi

Hasil perbandingan ditampilkan dalam dua bahasa (Indonesia dan Inggris) yang mencakup:

- Skor Kemiripan (%)
- Label Kemiripan Kategorikal
- Ringkasan AI dalam enam bagian utama:
 - 1) Ringkasan Jurnal Pertama
 - 2) Ringkasan Jurnal Kedua
 - 3) Kesamaan Utama
 - 4) Perbedaan Utama
 - 5) Analisis Hubungan
 - 6) Kesimpulan

Gambar 15 menampilkan hasil perbandingan dalam Bahasa Inggris.



Gambar 15. Hasil komparasi dan ringkasan jurnal (Bahasa Inggris)

Selanjutnya Gambar 16 menampilkan hasil perbandingan dalam Bahasa Indonesia



Gambar 16. Hasil komparasi dan ringkasan jurnal (Bahasa Indonesia)

3. Fungsi dan Manfaat Sistem

Dengan adanya implementasi sistem ini, pengguna dapat:

- Melakukan perbandingan jurnal secara otomatis dan efisien.
- Mendapatkan interpretasi semantik dari kesamaan jurnal.
- Melihat hasil analisis berbasis AI dalam dua bahasa.
- Mengidentifikasi keterkaitan, perbedaan, dan relevansi antar dua karya ilmiah.

F. Pengujian Sistem

Pengujian sistem bertujuan untuk memastikan bahwa fungsionalitasnya berjalan sesuai dengan tujuan perancangan dan untuk mengukur efektivitasnya dalam menangani skenario penggunaan secara nyata. Pengujian ini dilakukan dengan pendekatan *Black-Box Testing*, yang berfokus pada evaluasi fungsionalitas dan kinerja sistem berdasarkan input yang diberikan dan output yang dihasilkan, tanpa memeriksa struktur kode internal. Pengujian mencakup berbagai fitur utama, seperti proses unggah jurnal, penentuan skor kemiripan, klasifikasi label, hingga tampilan hasil ringkasan jurnal secara bilingual. Hasil pengujian fungsional dirangkum pada Tabel 5 dan Tabel 6.

Tabel 5. Hasil Pengujian Sistem dengan Metode *Blackbox Testing*

No	Test Case	Skenario Pengujian	Hasil yang di Harapkan	Hasil
1	Unggah Dua Jurnal PDF	Pengguna mengklik tombol “Go to Compare” dari halaman beranda	Sistem dapat menampilkan pratinjau isi kedua jurnal	[v] Berhasil [] Gagal
2	Validasi unggahan dua jurnal	Pengguna mencoba mengunggah lebih dari dua jurnal	Sistem menolak unggahan dan menampilkan pesan kesalahan	[v] Berhasil [] Gagal
3	Respons sistem terhadap teks kosong	Pengguna mengunggah jurnal dengan abstrak kosong atau rusak	Sistem memberikan notifikasi kesalahan input	[v] Berhasil [] Gagal
4	Proses Komparasi Jurnal	Sistem memproses kemiripan semantik berdasarkan abstrak dari dua jurnal yang telah diunggah	Sistem menampilkan skor kemiripan dalam persentase serta label kemiripan kategori	[v] Berhasil [] Gagal

5	Hasil Skor Kemiripan Rendah	Sistem diuji dengan dua jurnal berbeda topik: kanker dan peramalan air	Sistem menampilkan skor kemiripan rendah ($\approx 10\%$) dan label “Tidak Relevan”	[v] Berhasil [] Gagal
6	Hasil Skor Kemiripan Menengah	Sistem diuji dengan jurnal terkait sistem rekomendasi dan <i>black-box testing</i>	Sistem menampilkan skor sekitar 19% dan label “Sedikit Berkaitan”	[v] Berhasil [] Gagal
7	Hasil Skor Kemiripan Cukup Terkait	Sistem diuji dengan dua jurnal yang membahas model embedding NLP	Sistem menampilkan skor sekitar 30% dan label “Cukup Berkaitan”	[v] Berhasil [] Gagal
8	Hasil Skor Kemiripan Tinggi	Sistem diuji dengan dua jurnal yang membahas pengelompokan dokumen menggunakan BERT	Sistem menampilkan skor tinggi ($\approx 75\%$) dan label “Sangat Berkaitan”	[v] Berhasil [] Gagal
6	Tampilan Bahasa AI <i>Narrative</i> (EN & ID)	Pengguna memilih mode tampilan narasi perbandingan hasil dalam bahasa Inggris dan Bahasa Indonesia	Sistem menampilkan narasi perbandingan AI dalam dua bahasa (EN & ID) berdasarkan hasil analisis kedua jurnal	[v] Berhasil [] Gagal
7	Respons Sistem terhadap <i>Reset Upload</i>	Pengguna mengklik tombol “ <i>Reset Upload</i> ” setelah mengunggah dua jurnal	Sistem menghapus tampilan jurnal sebelumnya dan kembali ke status awal (Halaman Beranda)	[v] Berhasil [] Gagal

Setelah dilakukan serangkaian pengujian sebagaimana dirangkum pada Tabel 5, tahap selanjutnya adalah mengevaluasi tingkat akurasi sistem dalam menentukan skor dan label kemiripan semantik antar jurnal. Tabel 6 berikut menyajikan evaluasi tingkat akurasi skor dan label kemiripan semantik, sekaligus mengintegrasikan metrik kinerja berupa waktu respons untuk setiap skenario pengujian:

Tabel 6. Evaluasi Akurasi Fungsional dan Kinerja Sistem

Test Case	Skenario Pengujian	Skor	Label	Waktu Respon (Detik)	Status
Uji Jurnal Tidak Relevan	Menguji dua jurnal dari topik yang sama sekali berbeda (kanker dan ekosistem air).	9.00%	Tidak Relevan	14.73	[v] Berhasil [] Gagal
Uji Jurnal Sedikit Berkaitan	Menguji pasangan jurnal yang memiliki hubungan lemah atau konteks berbeda dalam domain, namun masih dalam lingkup umum teknologi.	19.00%	Sedikit Berkaitan	15.08	[v] Berhasil [] Gagal
Uji Jurnal Cukup Berkaitan	Menguji dua jurnal bertopik teknologi informasi yang menerapkan <i>sentence embedding</i> dan teknik klasifikasi berbasis <i>vector</i> .	30.00%	Cukup Berkaitan	13.15	[v] Berhasil [] Gagal
Uji Jurnal Sangat Berkaitan	Menguji kesamaan semantik antara dua jurnal dengan topik dan pendekatan yang sangat serupa.	75.00%	Sangat Berkaitan	18.03	[v] Berhasil [] Gagal

Hasil pada Tabel 6 menunjukkan bahwa sistem mampu melakukan klasifikasi terhadap berbagai tingkat kemiripan abstrak jurnal. Skor kemiripan yang ditampilkan juga memiliki koherensi terhadap konteks semantik dari masing-masing pasangan jurnal. Hal ini membuktikan bahwa integrasi model *Sentence-Transformers* dengan algoritma *K-Means* dan penetapan label berbasis *centroid* mampu memberikan hasil yang representatif dalam konteks analisis komparatif jurnal ilmiah.

Dari sisi kinerja, data waktu respons memberikan bukti kuantitatif mengenai efektivitas sistem dalam membantu peneliti. Ringkasan statistik kinerja adalah sebagai berikut:

- Waktu respons rata-rata: 15.25 detik
- Waktu respons tercepat: 13.15 detik
- Waktu respons terlama: 18.03 detik

Dengan waktu proses rata-rata hanya sekitar 15 detik untuk menganalisis dan menyajikan perbandingan dua jurnal, sistem ini terbukti jauh lebih efisien dibandingkan proses analisis manual. Proses manual yang melibatkan pembacaan dan sintesis dua jurnal dapat memakan waktu beberapa menit. Dengan demikian, sistem ini secara efektif dapat memangkas waktu analisis dan mempercepat alur kerja tinjauan literatur bagi peneliti.

BAB V

KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini dilakukan untuk menjawab tantangan dalam membandingkan jurnal ilmiah berdasarkan kemiripan semantik, dengan memanfaatkan pendekatan modern dalam pemrosesan bahasa alami. Sistem yang dirancang menggabungkan model *Sentence-Transformers* dan algoritma *K-Means Clustering* dapat digunakan untuk mengelompokkan dan membandingkan jurnal berdasarkan abstraknya secara otomatis. Dari hasil implementasi dan pengujian sistem, diperoleh beberapa kesimpulan berikut:

1. Penelitian ini berhasil menerapkan *Sentence-Transformers* menggunakan model *all-mpnet-base-v2* untuk menghitung skor kemiripan semantik antar abstrak jurnal. Skor yang dihasilkan kemudian dikelompokkan menggunakan algoritma *K-Means* menjadi empat kategori: Tidak Relevan, Sedikit Berkaitan, Cukup Berkaitan, dan Sangat Berkaitan. Kombinasi ini memungkinkan sistem untuk melakukan pemetaan hubungan antar jurnal secara terstruktur dan berbasis konteks, sehingga mempermudah analisis kemiripan di antara dokumen ilmiah.
2. Sistem yang dibangun terbukti efektif dalam membantu peneliti membandingkan jurnal secara semantik. Melalui serangkaian pengujian, sistem mampu menghasilkan skor kemiripan dan label kategori yang sesuai dengan isi dan topik jurnal yang dibandingkan. Lebih lanjut, efektivitas sistem ini terbukti secara kuantitatif melalui pengujian kinerja, di mana sistem mampu menyelesaikan proses perbandingan secara otomatis dengan waktu komputasi rata-rata 15.25 detik. Dengan kecepatan proses ini, ditambah dengan fitur narasi otomatis dalam dua bahasa, secara signifikan memangkas waktu yang dibutuhkan untuk menilai keterkaitan antar jurnal dibandingkan dengan metode manual, sehingga hasil komparasi menjadi lebih mudah dipahami baik secara teknis maupun konseptual.

B. Saran

Penelitian ini telah berhasil mengimplementasikan sistem untuk perbandingan semantik antara dua jurnal secara spesifik. Untuk pengembangan selanjutnya, disarankan agar sistem ini ditransformasi dengan menambahkan fitur perbandingan satu banding banyak (*one-vs-many*), yang memungkinkan pengguna cukup mengunggah satu jurnal utama untuk kemudian dibandingkan secara otomatis dengan korpus literatur yang luas dari basis data akademik daring seperti *Semantic Scholar* atau *Google Scholar*. Dengan menghasilkan daftar jurnal yang diurutkan berdasarkan tingkat relevansi semantiknya, fungsionalitas sistem akan berevolusi dari alat komparasi menjadi sebuah mesin eksplorasi dan rekomendasi literatur yang proaktif. Hal ini akan secara signifikan mempermudah dan mempercepat proses tinjauan pustaka serta membantu peneliti menemukan referensi-referensi kunci yang paling relevan dengan topik kajian mereka secara lebih efisien.

DAFTAR PUSTAKA

- Afriyani, S., Surono, S., & Solihin, M. I. (2024). *Chi-Square Feature Selection with Pseudo-Labeling in Natural Language Processing*. 8(3), 896–909.
- Anggrainingsih, R., Wihidayat, E. S., & Widoyono, B. (2024). Sentence embedding to improve rumour detection performance model. *IAES International Journal of Artificial Intelligence*, 13(1), 115–121. <https://doi.org/10.11591/ijai.v13.i1.pp115-121>
- Annisa, R., Rosiyadi, D., & Riana, D. (2020). Improved point center algorithm for k-means clustering to increase software defect prediction. *International Journal of Advances in Intelligent Informatics*, 6(3), 328–339. <https://doi.org/10.26555/ijain.v6i3.484>
- Aszani, A., Wicaksono, H. I., Nadzima, U., & Heryawan, L. (2023). Information Retrieval for Early Detection of Disease Using Semantic Similarity. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 17(1), 45. <https://doi.org/10.22146/ijccs.80077>
- Ayuningtyas, P. K., Atmodjo WP, D., & Rachmadi, P. (2023). Performance And Functional Testing With The Black Box Testing Method. *International Journal of Progressive Sciences and Technologies*, 39(2), 212. <https://doi.org/10.52155/ijpsat.v39.2.5471>
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1). <https://doi.org/10.1057/s41599-021-00903-w>
- Chikkarangaiah, J., Uday, A., De Hilster, D., Gangadhar, S., & Shetty, J. (2024). Enhancing the English natural language processing dictionary using natural language processing. *IAES International Journal of Artificial Intelligence*, 13(3), 3466–3477. <https://doi.org/10.11591/ijai.v13.i3.pp3466-3477>
- Colangelo, M. T., Meleti, M., Guizzardi, S., Calciolari, E., & Galli, C. (2025). A

- Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata. *Big Data and Cognitive Computing*, 9(3), 1–18. <https://doi.org/10.3390/bdcc9030067>
- Desta Yolanda, Mohammad Hafiz Hersyah, & Eno Marozi. (2021). Implementasi Metode Unsupervised Learning Pada Sistem Keamanan Dengan Optimalisasi Penyimpanan Kamera IP. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(6), 1099–1105. <https://doi.org/10.29207/resti.v5i6.3552>
- Hambi, E. M., & Benabbou, F. (2020). A deep learning based technique for plagiarism detection: A comparative study. *IAES International Journal of Artificial Intelligence*, 9(1), 81–90. <https://doi.org/10.11591/ijai.v9.i1.pp81-90>
- Harahap, M., Zamili, A. W. D. R., Arvansyah, M. A., Saragih, E. F., Rajen, S., & Husein, A. M. (2022). K-Means Clustering Algorithm Approach in Clustering Data on Cocoa Production Results in the Sumatra Region. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(6), 905–910. <https://doi.org/10.29207/resti.v6i6.4199>
- Irbah salsabila, & Yuliant Sibaroni. (2021). Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(3), 520–526. <https://doi.org/10.29207/resti.v5i3.3078>
- Jia, J., Liang, W., & Liang, Y. (2023). *A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing*. 1–23. <http://arxiv.org/abs/2312.05589>
- Kurnianingsih, Wirasatriya, A., Lazuardi, L., Wibowo, A., Pradekso, B. K., Prasetyo, S., Aji, N. B., & Sato-Shimokawara, E. (2023). Transformer-based Deep Learning for COVID-19 Prediction Based on Climate Variables in Indonesia. *International Journal on Advanced Science, Engineering and Information Technology*, 13(2), 632–637. <https://doi.org/10.18517/ijaseit.13.2.18292>

- Kusumaningrum, R., Khoerunnisa, S. F., Khadijah, K., & Syafrudin, M. (2024). Exploring Community Awareness of Mangrove Ecosystem Preservation through Sentence-BERT and K-Means Clustering. *Information (Switzerland)*, 15(3), 1–14. <https://doi.org/10.3390/info15030165>
- Octavian Ery Pamungkas, Puspa Rahmawati, Dhany Maulana Supriadi, Natasya Nur Khalika, Thofan Maliyano, Dicky Revan Pangestu, Nugraha, E. S., Mas Aly Afandi, Nurcahyani Wulandari, Petrus Kerowe Goran, & Agung Wicaksono1. (2022). Classification of Rupiah to Help Blind with The Convolutional Neural Network Method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(2), 259–268. <https://doi.org/10.29207/resti.v6i2.3852>
- Ortakci, Y. (2024). Revolutionary text clustering: Investigating transfer learning capacity of SBERT models through pooling techniques. *Engineering Science and Technology, an International Journal*, 55(December 2023), 101730. <https://doi.org/10.1016/j.jestch.2024.101730>
- Rustam, Z., Hartini, S., Novkaniza, F., Pandelaki, J., Hidayat, R., & Ezziyyani, M. (2023). An Efficient and Robust Ischemic Stroke Detection Using a Combination of Convolutional Neural Network (CNN) and Kernel KMeans Clustering. *International Journal on Advanced Science, Engineering and Information Technology*, 13(3), 969–974. <https://doi.org/10.18517/ijaseit.13.3.18264>
- Seo, D., Kim, S., Oh, S., & Kim, S. H. (2022). K-Means Clustering-Based Safety System in Large-Scale Industrial Site Using Industrial Wireless Sensor Networks. *Sensors*, 22(8), 1–14. <https://doi.org/10.3390/s22082897>
- Simanjuntak, H. T. A., Silaban, P. E. P., Manurung, J. K. S., & Sormin, V. H. (2023). Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 10(3), 641–652. <https://doi.org/10.25126/jtiik.20231026468>
- Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data*, 9(1).

<https://doi.org/10.1186/s40537-022-00564-9>

Sy, C. Y., Maceda, L. L., & Abisado, M. B. (2024). AI-driven analysis: optimizing tertiary education policy through machine learning insights. *International Journal of Advances in Intelligent Informatics*, 10(2), 296–316. <https://doi.org/10.26555/ijain.v10i2.1525>

Van Son, N., Huong, L. T., & Thanh, N. C. (2021). A two-phase plagiarism detection system based on multi-layer lstm networks. *IAES International Journal of Artificial Intelligence*, 10(3), 636–648. <https://doi.org/10.11591/ijai.v10.i3.pp636-648>

Weng, M. H., Wu, S., & Dyer, M. (2022). Identification and Visualization of Key Topics in Scientific Publications with Transformer-Based Language Models and Document Clustering Methods. *Applied Sciences (Switzerland)*, 12(21). <https://doi.org/10.3390/app122111220>

Wong, B., & Tanaka, K. (2025). *High-Fidelity Pseudo-label Generation by Large Language Models for Training Robust Radiology Report Classifiers*. 1–9. <http://arxiv.org/abs/2505.01693>

Yee, G. P., Rusiman, M. S., Ismail, S., Suparman, Hamzah, F. M., & Shafi, M. A. (2023). K-means clustering analysis and multiple linear regression model on household income in Malaysia. *IAES International Journal of Artificial Intelligence*, 12(2), 731–738. <https://doi.org/10.11591/ijai.v12.i2.pp731-738>

Zainuddin, Z., & Risal, A. A. N. (2024). Balanced clustering for student admission school zoning by parameter tuning of constrained k-means. *IAES International Journal of Artificial Intelligence*, 13(2), 2299–2311. <https://doi.org/10.11591/ijai.v13.i2.pp2301-2313>

LAMPIRAN

Lampiran 1. Dataset Abstrak Jurnal

A	B	C	D
title1	abstract1	title2	abstract2
The Health Utilities Index (HUI): concep	This is a review of the Health Utilities I	Perceived Usefulness, Ease of Use, an	This paper presents the findings of two
Internet paradox. A social technology t	The Internet could change the lives of	Ecological momentary interventions: i	PURPOSE Psychosocial and health beha
Acculturation and Latino health in the	This chapter provides an overview of t	Dissecting racial bias in an algorithm u	Racial bias in health algorithms The U.S
Best Practices for Developing and Valid	Scale development and validation are	'Old Economy' Inputs for 'New Econo	This paper discusses the results of a tw
Contamination of drinking-water by ar	The contamination of groundwater by	Linear Economy Versus Circular Econo	Abstract Upon visiting the existing liter
Collaborative Economy and Tourism: C	House swapping, ridesharing, volunt	Technology development for the prod	A biorefinery that supplements its mar
Toward a Greater Understanding of Ent	The informal economy consists of busi	Evaluating the public health impact of	Progress in public health and communi
Shaping Agility through Digital Options	Agility is vital to the innovation and co	Police Brutality and Black Health: Setti	We investigated links between police
Environment, economy and society: fit	Sustainable development is a conteste	Electronic Data Interchange and Small	Many EDI researchers and practitioners
Can electronic medical record systems	To broadly examine the potential heal	The world economy : historical statisti	Angus Maddison provides a comprehen
The Relationship between Social Cohes	Social cohesion involves the interspers	Technology as an occasion for structur	New medical imaging devices, such as
The health benefits of urban green spa	BACKGROUND Urban development pro	The Duality of Technology: Rethinking	This paper develops a new theoretical
Perceived Usefulness, Perceived Ease	(Valid measurement scales for predicti	Chains and networks, territories and s	A vast and continually expanding litera
What Do We Learn from the Weather?	A rapidly growing body of research app	Hispanic health in the USA: a scoping r	Hispanics are the largest minority grou
What the evidence shows about patien	Patient engagement is an increasingly	Mental Health Smartphone Apps: Revi	Background The number of mental hea
Trasformismo or transformation? The g	Abstract What does IPE have to contrib	Photovoice: a participatory action rese	Photovoice is a participatory action res
The Impact of Knowledge-Based Econo	The aim of this paper is to examine the	Capturing the Complexity in Advances	The past decade has brought advanced
Socioeconomic disparities in health in	OBJECTIVES We aimed to describe soci	Does health literacy mediate the relat	While socioeconomic disparities are an
Rethinking the Informal Economy: Link	This paper explores the relationship of	Antecedents of trust in the sharing ec	Users and potential users of the sharin
Market, Hierarchy, and Trust: The Know	Recent conceptualizations of trends in	The COVID-19 Pandemic and Its Impac	Abstract The COVID-19 pandemic broke
Globalization and history : the evolutio	Globalization is not a new phenomeno	Systematic review of the relationships	BackgroundGiven the rapid developme
User Acceptance of Computer Technolo	Computer systems cannot improve org	Mental health problems, use of ment	CONTEXT The US military has conducte
Factors Associated With Mental Health	Key Points Question What factors are a	Globalization, the new economy, and	The globalized new economy is bound
Review: Mental health impacts of the	Background: The COVID-19 pandemic h	How Chiefs Come to Power: The Politi	1. Introduction: the nature of political
The problem with the phrase women a	Intersectionality is a theoretical frame	Health behavior and health education	Health behavior change is our greatest

A	B	C	D
Digital Economy, Technological Innovat	Technological innovation and high-qua	Mobile text messaging for health: a sy	The aim of this systematic review of re
The Gifts Of Athena Historical Origins	CThank you for reading the gifts of athe	Unequal treatment: confronting racial	Racial and ethnic disparities in health c
Designing the Business Models for Circ	Switching from the current linear mod	The Global Platform Economy: A New	Global online platforms match firms wi
National Health Care Spending In 2020: US	health care spending increased 9.7	The order of economic liberalization :	Can knowledge of financial policies in c
An Inquiry into the Nature and Causes	It was the publication in 1776 of Adam	Information rules - a strategic guide to	From the Publisher: Information Goods
The New Global Economy and Develop	Policy makers in the developing world	Mental Health: Culture, Race, and Ethr	Mental health is fundamental to health
Implicit Racial/Ethnic Bias Among Healt	BACKGROUND In the United States, pe	Community health workers in low-, m	Over the past half-century, community
Integrating Data On Social Determinant	As population health becomes more o	A Conceptual and Operational Definiti	The acceptance of new information tec
Information Technology Adoption Acro	The process of information technology	US Department of Health and Human	More than ever before the US Departm
The Impact of COVID-19 on Health Beh	The COVID-19 pandemic has disrupted	Indoor air pollution in developing cou	Around 50% of people, almost all in de
Location, Competition, and Economic	Economic geography during an era of	Time Flies When You're Having Fun: C	Extant explanations of why users beha
Applying the Technology Acceptance M	In this study, we consider the online c	Understanding and Responding to Hea	Evidence of a social gradient in health
Why cultural safety rather than cultural	BackgroundEliminating indigenous anc	National Health Care Spending In 2021	Health care spending in the US grew 2.
A Resource-Based Perspective on Infor	The resource-based view of the firm a	You Say Illegal, I Say Legitimate: Entre	The entrepreneurial process drives eco
Predicting User Intentions: Comparing	Information systems IS cannot be effe	Extreme Weather and Climate Change	Extreme weather and climate events, s
The American Political Economy: Institi	Part I: Making Sense of the Political Ec	Health literacy in Europe: comparative	Background: Health literacy concerns th
Information Technology, Workplace Or	Recently, the relative demand for skill	Economic Effects of Coronavirus Outbr	This report discusses the economic imp
Stress and the Mental Health of Popula	This article provides an overview of re	A New Dimension of Health Care: Syst	Background There is currently a lack of
Health literacy as a public health goal: :	SUMMARY Health literacy is a relativel	Chemical Pesticides and Human Healt	The industrialization of the agricultural
Health complaints, stress, and distress:	Most current models in health psychol	The influence of education on health:	Background A clear understanding of th
Social Cognitive Theory and Individual	A model, based on Bandura's Social Co	The grounded psychometric developm	BackgroundHealth literacy has become
Loneliness as a public health issue: the	OBJECTIVES We aimed to determine w	The role of behavioral science theory	Increasing evidence suggests that publi
Methods for the estimation of the Nati	BACKGROUND Cost-effectiveness anal	Measuring health literacy in populatio	BackgroundSeveral measurement tools
Colonial global economy: towards a the	Abstract Standard accounts of the eme	Globalization in Question. The Interna	Introduction: globalization - a necessar
A conceptual framework for action on t	When he announced his intention to c	The outbreak of COVID-19 coronavirus	Background: The current outbreak of C

A	B	C	D
The Relationship Between Education and Health	Adults with higher educational attainment	Associations of health literacy with socioeconomics	Background Health literacy concerns the ability to understand and use basic health information to make decisions about one's health
On the Economy of the Human Process	Abstract : An approach to human performance	Prioritizing the Mental Health and Well-being	The COVID-19 pandemic has had an unprecedented impact on mental health
Understanding associations among race and health	Race/ethnicity and socioeconomic status	Review: Information Technology and Health	Despite the importance to researchers, clinicians, and patients, the literature on the topic is fragmented
The Sustainability of Evidence-Based Interventions	There is strong interest in implementing evidence-based interventions	Governing the economy : the politics of health	For over one hundred years, the British government has been involved in the regulation of the economy
Mobile medical and health apps: state of the field	This paper examines the state of the field	The prescription opioid and heroin crisis	Public health authorities have described the crisis as a public health emergency
Governance in the Blockchain Economy	Blockchain technology is often referred to as the 'distributed ledger'	Managerial Ties and Firm Performance	Using survey data from China, we demonstrate that managerial ties are positively related to firm performance
Social isolation, loneliness and health	The health and well-being consequences of social isolation and loneliness	The GRADE Evidence to Decision (EtD) Framework	Objective To describe a framework for the development of EtD statements
Health Care Pollution And Public Health	An up-to-date assessment of environmental health	Privacy in Context: Technology, Policy, and Practice	Privacy is one of the most urgent issues in health care
The Political Economy of Terrorism: Theoretical and Empirical Perspectives	The Political Economy of Terrorism	The Role of Telehealth in Reducing the Burden of Disease	Xiaoyun Zhou, MMH Centaine L. Snosw
The global burden of oral diseases and conditions	This paper outlines the burden of oral diseases and conditions	Differences between caregivers and non-caregivers	Providing care for a frail older adult has become a challenge for many families
What's Driving the New Economy?: The Role of Technology	Using a unique nationally representative survey	The World Oral Health Report 2003: Conclusions and Recommendations	Chronic diseases and injuries are the leading causes of disability and death
Valuing health-related quality of life: A new version of the EQ-5D	A new version of the EQ-5D, the EQ-5D-5L	Circular Economy and Sustainability of Textiles	are essential to humans in a value chain
Development of an Instrument to Measure Health Literacy	This paper reports on the development of the instrument	Digital Mental Health and COVID-19: U.S. Interest in and Use of Telehealth During the Pandemic	As interest in and use of telehealth during the pandemic have increased
A meta-analysis of self-determination theory	ABSTRACT There are no literature reviews on self-determination theory	The Mismanagement of Talent: Employment and the Knowledge Economy	The knowledge economy conjures a vision of a new economy
The SEED and the Rapid Annotation of Tool Development	In 2004, the SEED (http://pubseed.theseed.org) was developed	Energy structure, digital economy, and the environment	As a new production factor, digitalization is becoming an important driving force
The mental health continuum: from mental illness to mental well-being	This paper introduces and applies an approach to the mental health continuum	Using Electronic Health Records for Population Health Research	The use and functionality of electronic health records for population health research
Measuring social class in US public health research	Increasing social inequalities in health	Knowledge of the Firm, Combinative Capabilities, and the New Product Development Process	How should we understand why firms innovate?
What is the meaning of health literacy?	The objective of this review was to clarify the meaning of health literacy	Stigma, mental health, and resilience	OBJECTIVES We assessed the associations between stigma, mental health, and resilience
Blockchain technology and its relationship to the environment	Globalisation of supply chains makes the environment more vulnerable	Mental health literacy: empowering the workforce	For major physical diseases, it is widely recognized that mental health literacy is an important determinant
Atom Economy-A Challenge for Organic Synthesis	Enhancing the efficiency of the synthesis	Review of community-based research in public health	Community-based research in public health is a valuable approach to understanding the social determinants of health
Climate change and mental health: risk factors and protective factors	Background This article provides an overview of the current state of knowledge	The benefits of health information technology	An unprecedented federal effort is underway to improve the health of the nation
Principles of Political Economy with Special Reference to the History of the Theory	Get Principles Of Political Economy With Special Reference To The History Of The Theory	The rise of the "just-in-time" workforce	The so-called "gig-economy" has been growing rapidly
Grundrisse : foundations of the critique of political economy	Written during the winter of 1857-8, this work is the first of three volumes	Using PLS path modeling in new technology development	Purpose - Partial least squares (PLS) path modeling is a statistical technique
Evidence suggesting that a chronic disease is a risk factor for mental health	OBJECTIVES This study evaluated the effectiveness of the new tools	Tools in Comparative Political Economy	This article introduces a large new cross-national dataset
A Meta-Analysis of the Effectiveness of The Health Belief Model (HBM)	Rosens et al. The Health Belief Model (HBM)	The Broad Scope of Health Effects from Environmental Exposures	Background: Concerns for arsenic exposure in drinking water

A	B	C	D
Technology Infrastructure and Business Model Innovation	The general objective of the study was to investigate the factors influencing technology acceptance	Technology Acceptance Model 3 and a Prior research has provided valuable information	Prior research has provided valuable information on the factors influencing technology acceptance
A framework for public health action: the 5-tier pyramid	A 5-tier pyramid best describes the impact of public health action	Exposure to Neighborhood Green Space and Mental Health	Green space is now widely viewed as a public health asset
Sleep Health: An Opportunity for Public Health Action	The concept of sleep health provides a framework for public health action	The novel coronavirus (COVID-2019) outbreak	The 2019 novel coronavirus (COVID-2019) outbreak has caused a global health crisis
Evaluating Strategies For Reducing Health Inequalities	The opportunities for healthy choices in the food system	Implementation of Circular Economy in Small and Medium-sized Enterprises (SMEs)	Small and medium-sized enterprises (SMEs) are the backbone of the economy
Monetary Policy and Exchange Rate Volatility	We lay out a small open economy version of the monetary policy	Effects of soft drink consumption on obesity	In a meta-analysis of 88 studies, we examined the effects of soft drink consumption on obesity
The effectiveness of M-health technologies in improving health outcomes	Background The application of mobile health technologies in improving health outcomes	THE MORAL ECONOMY OF THE ENGLISH WE HAVE BEEN WARNED IN RECENT YEARS	THE MORAL ECONOMY OF THE ENGLISH WE HAVE BEEN WARNED IN RECENT YEARS
User Acceptance of Information Technology	Abstract Lack of user acceptance has been a major barrier to the widespread use of information technology	Broad strokes towards a grand theory of the firm	ABSTRACT This article aims to provide a broad strokes towards a grand theory of the firm
The Political Economy of Government Intervention	The determinants of government intervention in the health care market	Product design and business model strategy	The transition within business from a linear to a circular model
Enhancing treatment fidelity in health care	Treatment fidelity refers to the methodological rigor of the intervention	Consumer Acceptance and Use of Information Technology	This paper extends the unified theory of acceptance and use of technology
Mental illness and/or mental health? In a continuous assessment and a categorical approach	A theoretical extension of the Technology Acceptance Model	The present research develops and tests a theoretical model	The present research develops and tests a theoretical model
Qualitative data analysis for health services research	OBJECTIVE To provide practical strategies for qualitative data analysis	Circular Economy - Challenges for the Future	Abstract The circular economy model has gained significant attention
Self-rated health and mortality: a review	We examine the growing number of studies on self-rated health and mortality	The National Institute on Minority Health and Health Disparities	We introduce the National Institute on Minority Health and Health Disparities
Social and Structural Determinants of Health	Since the World Health Organization launched the concept of the social determinants of health	When patient activation levels change	Patient engagement has become a major challenge for health care providers
Compassion and Repression: The Moral Dimensions of Immigration Policies	Immigration policies in Europe in the 21st century	Consumer health information seeking	Increasingly, consumers engage in health information seeking
URBAN REGIMES AND THE CAPACITY TO ADAPT TO CLIMATE CHANGE	ABSTRACT: Regime theory starts with the idea of a dominant regime	Discrete choice experiments in health economics	Discrete choice experiments (DCEs) have become a popular method in health economics
The Structure and Performance of Financial Markets	The paper examines the determinants of the structure and performance of financial markets	The prevalence, distribution, and management of mental health disorders	The survey data presented here are on the prevalence, distribution, and management of mental health disorders
The Platform Political Economy of FinTech	ABSTRACT 'FinTech' is the digital sector of the financial industry	Patient-centred access to health care: Background	Access is central to the performance of health care
Integration of Trade and Disintegration of the Firm	The last few decades have seen a specific pattern of trade and disintegration	The Role of the Business Model in Capital Structure	This paper explores the role of the business model in capital structure
Mental Health of Children and Adolescents	Background: The COVID-19 pandemic has had a significant impact on the mental health of children and adolescents	Bayesian Estimation of an Open Economy	In this paper we develop a dynamic stochastic general equilibrium model
Towards a Circular Economy: The Role of the Government	While there is great potential in the circular economy, there are several barriers	Barriers and facilitators to the use of health information technology	Limited attention has been paid to how barriers and facilitators to the use of health information technology
Validation and utility of a self-report of health status	CONTEXT The Primary Care Evaluation of User Acceptance and Use of Information Technology	The Globalization of Liberalization: Political and Economic Context	One of the most important developments in the world is the globalization of liberalization
Meta-analysis of the relationship between health promotion programs and health outcomes	BACKGROUND Risk perceptions are central to health promotion programs	Optimism, coping, and health: assessment and intervention	This article describes a scale measuring optimism, coping, and health
Translating Social Ecological Theory into Health Promotion Programs	Health promotion programs often lack a theoretical basis	Challenges and burden of the Coronavirus disease 2019 (COVID-19)	The coronavirus disease 2019 (COVID-19) has caused a global health crisis
Farewell to Socialism@@@The Social Contract	This book provides a comprehensive analysis of the social contract	The Long-Term Labor Market Consequences of the COVID-19 Pandemic	This paper studies the labor market consequences of the COVID-19 pandemic
The Effect of Service Employees' Technological Skills on Their Performance	The goal of the present paper is to identify the factors influencing the performance of service employees	Standards of Care for the Health of Transgender and Gender-Nonconforming People	Abstract Background: Transgender and gender-nonconforming people

Lampiran 2. Dataset Setelah *Pseudo-Labeling*

A	B	C	D	E
title1	abstract1	title2	abstract2	pseudo-label
The Health Utilities Index (HUI): This is a review of the Health Utility Index (HUI) and its use in health economics.	Perceived Usefulness, Ease of Use, and Usefulness: This paper presents the findings of a study on the perceived usefulness, ease of use, and usefulness of a health utility index (HUI) in health economics.	Ecological momentary intervention (EMI) for smoking cessation: This paper presents the findings of a study on the effectiveness of an ecological momentary intervention (EMI) for smoking cessation.	PURPOSE Psychosocial and health-related quality of life (HRQL) outcomes of a randomized controlled trial (RCT) of a biorefinery that supplements the production of bioethanol with the production of biogas.	0.180239886
Internet paradox. A social technology that could change the lives of people in the developing world.	The Internet could change the lives of people in the developing world by providing access to information and services that were previously unavailable.	Dissecting racial bias in an algorithm: Race, class, and gender in health algorithms.	Racial bias in health algorithms: This paper discusses the results of a study on racial bias in health algorithms.	0.391628027
Best Practices for Developing an eHealth System: A review of the literature on best practices for developing an eHealth system.	Scale development and validation of a measure of eHealth literacy: This paper discusses the results of a study on the development and validation of a measure of eHealth literacy.	'Old Economy' Inputs for 'New Economy' Outputs: A review of the literature on the relationship between the 'old economy' and the 'new economy'.	This paper discusses the results of a study on the relationship between the 'old economy' and the 'new economy'.	0.371246964
Contamination of drinking-water in the developing world: A review of the literature on the contamination of drinking-water in the developing world.	The contamination of groundwater in the developing world: This paper discusses the results of a study on the contamination of groundwater in the developing world.	Linear Economy Versus Circular Economy: A review of the literature on the relationship between the linear economy and the circular economy.	Abstract Upon visiting the existing literature, the authors found that the linear economy is a more sustainable model than the circular economy.	0.054675169
Collaborative Economy and Tourism: A review of the literature on the collaborative economy and tourism.	House swapping, ridesharing, and other collaborative economy activities: This paper discusses the results of a study on the collaborative economy and tourism.	Technology development for the development of a biorefinery that supplements the production of bioethanol with the production of biogas.	A biorefinery that supplements the production of bioethanol with the production of biogas.	0.10084822
Toward a Greater Understanding of the Informal Economy: A review of the literature on the informal economy.	The informal economy consists of a large number of small, unregistered, and often illegal businesses that operate outside the formal economy.	Evaluating the public health impact of a biorefinery that supplements the production of bioethanol with the production of biogas.	Progress in public health and community development: This paper discusses the results of a study on the progress in public health and community development.	-0.016242785
Shaping Agility through Digital Capabilities: A review of the literature on digital capabilities and agility.	Agility is vital to the innovation and growth of a business in the digital age.	Police Brutality and Black Health: A review of the literature on police brutality and Black health.	We investigated links between police brutality and Black health.	0.026338331
Environment, economy and society: A review of the literature on the relationship between the environment, economy, and society.	Sustainable development is a concept that seeks to balance the needs of the present with the needs of the future.	Electronic Data Interchange and the Development of a Biorefinery: A review of the literature on electronic data interchange and the development of a biorefinery.	Many EDI researchers and practitioners have focused on the development of a biorefinery.	0.030276757
Can electronic medical records improve patient care? A review of the literature on electronic medical records and patient care.	To broadly examine the potential of electronic medical records to improve patient care.	The world economy: historical and contemporary perspectives: A review of the literature on the world economy.	Angus Maddison provides a comprehensive overview of the world economy.	0.139270499
The Relationship between Social Cohesion and the Informal Economy: A review of the literature on social cohesion and the informal economy.	Social cohesion involves the integration of individuals and groups into a community.	Technology as an occasion for social change: A review of the literature on technology and social change.	New medical imaging devices, such as magnetic resonance imaging (MRI) and computed tomography (CT) scans, have revolutionized medical diagnosis.	0.077206388
The health benefits of urban green spaces: A review of the literature on the health benefits of urban green spaces.	BACKGROUND Urban development and the health benefits of urban green spaces: This paper discusses the results of a study on the health benefits of urban green spaces.	The Duality of Technology: Rethinking the Informal Economy: A review of the literature on the duality of technology and the informal economy.	This paper develops a new theoretical framework for understanding the relationship between technology and the informal economy.	0.134024233
Perceived Usefulness, Perceived Ease of Use, and Usefulness: A review of the literature on perceived usefulness, perceived ease of use, and usefulness.	Valid measurement scales for perceived usefulness, perceived ease of use, and usefulness: This paper discusses the results of a study on valid measurement scales for perceived usefulness, perceived ease of use, and usefulness.	Chains and networks, territories and the environment: A review of the literature on chains and networks, territories, and the environment.	A vast and continually expanding body of research on the relationship between chains and networks, territories, and the environment.	0.060369313
What Do We Learn from the World? A review of the literature on the world.	A rapidly growing body of research on the world: This paper discusses the results of a study on the world.	Hispanic health in the USA: a social and cultural perspective: A review of the literature on Hispanic health in the USA.	Hispanics are the largest minority group in the USA.	0.001342501
What the evidence shows about Patient engagement is an increasing trend: A review of the literature on patient engagement.	Mental Health Smartphone App: Background The number of mental health smartphone apps is increasing.	Photovoice: a participatory action research method: A review of the literature on photovoice.	Photovoice is a participatory action research method that uses photography to explore and communicate community issues.	0.099247411
Trasformismo or transformation? A review of the literature on transformation.	Abstract What does IPE have to do with transformation? This paper discusses the results of a study on transformation.	Photovoice: a participatory action research method: A review of the literature on photovoice.	Photovoice is a participatory action research method that uses photography to explore and communicate community issues.	0.372995257
The Impact of Knowledge-Based Systems on the Informal Economy: A review of the literature on knowledge-based systems and the informal economy.	The aim of this paper is to examine the impact of knowledge-based systems on the informal economy.	Capturing the Complexity in Advancing the Informal Economy: A review of the literature on capturing the complexity in advancing the informal economy.	The past decade has brought about significant changes in the informal economy.	0.038940355
Socioeconomic disparities in health: A review of the literature on socioeconomic disparities in health.	OBJECTIVES We aimed to describe the relationship between socioeconomic disparities and health.	Does health literacy mediate the relationship between socioeconomic disparities and health? A review of the literature on health literacy and socioeconomic disparities.	While socioeconomic disparities in health are well-documented, the role of health literacy in mediating these disparities is less clear.	0.208763599
Rethinking the Informal Economy: A review of the literature on the informal economy.	This paper explores the relationship between the informal economy and the formal economy.	Antecedents of trust in the sharing economy: A review of the literature on antecedents of trust in the sharing economy.	Users and potential users of the sharing economy: This paper discusses the results of a study on the antecedents of trust in the sharing economy.	0.753816783
Market, Hierarchy, and Trust: A review of the literature on market, hierarchy, and trust.	Recent conceptualizations of the informal economy: This paper discusses the results of a study on the informal economy.	The COVID-19 Pandemic and Its Impact on the Informal Economy: A review of the literature on the COVID-19 pandemic and its impact on the informal economy.	Abstract The COVID-19 pandemic has had a significant impact on the informal economy.	0.354566872
Globalization and history: A review of the literature on globalization and history.	Globalization is not a new phenomenon: This paper discusses the results of a study on globalization and history.	Systematic review of the relationship between globalization and health: A review of the literature on globalization and health.	Background Given the rapid development of globalization, it is important to understand its relationship with health.	0.299443871
User Acceptance of Computer Technology: A review of the literature on user acceptance of computer technology.	Computer systems cannot improve mental health problems, use of computer systems is not recommended: This paper discusses the results of a study on user acceptance of computer technology.	Mental health problems, use of computer systems: A review of the literature on mental health problems and the use of computer systems.	CONTEXT The US military has a long history of using computer systems to improve mental health.	-0.008931525
Factors Associated With Mental Health: A review of the literature on factors associated with mental health.	Key Points Question What factors are associated with mental health? This paper discusses the results of a study on factors associated with mental health.	Globalization, the new economy, and the informal economy: A review of the literature on globalization, the new economy, and the informal economy.	The globalized new economy is a result of globalization.	0.044408891
Review: Mental health impacts (Background: The COVID-19 pandemic and its impact on mental health): A review of the literature on mental health impacts.	The COVID-19 pandemic and its impact on mental health: This paper discusses the results of a study on mental health impacts.	How Chiefs Come to Power: The 1. Introduction: the nature of power: A review of the literature on how chiefs come to power.	1. Introduction: the nature of power: This paper discusses the results of a study on how chiefs come to power.	0.061863072
The problem with the phrase 'intersectionality' is a theoretical issue: A review of the literature on intersectionality.	Intersectionality is a theoretical issue: This paper discusses the results of a study on intersectionality.	Health behavior and health education: A review of the literature on health behavior and health education.	Health behavior change is our goal: This paper discusses the results of a study on health behavior change.	0.007964596
				0.435359538

A	B	C	D	E
Digital Economy, Technological Innovation and the Informal Economy: A review of the literature on the digital economy, technological innovation, and the informal economy.	Technological innovation and the informal economy: This paper discusses the results of a study on technological innovation and the informal economy.	Mobile text messaging for health promotion: A review of the literature on mobile text messaging for health promotion.	The aim of this systematic review is to examine the effectiveness of mobile text messaging for health promotion.	0.082428083
The Gifts Of Athena: Historical Origins of the Invention of the Gift: A review of the literature on the gifts of Athena.	Thank you for reading the gifts of Athena: This paper discusses the results of a study on the gifts of Athena.	Unequal treatment: confronting racial and ethnic disparities in health care: A review of the literature on unequal treatment.	Racial and ethnic disparities in health care: This paper discusses the results of a study on racial and ethnic disparities in health care.	-0.037103318
Designing the Business Models for the Informal Economy: A review of the literature on designing the business models for the informal economy.	Switching from the current linear economy to a circular economy: This paper discusses the results of a study on designing the business models for the informal economy.	The Global Platform Economy: A review of the literature on the global platform economy.	A Global online platforms match the needs of the informal economy.	0.342005253
National Health Care Spending in the United States: A review of the literature on national health care spending in the United States.	US health care spending increases: This paper discusses the results of a study on national health care spending in the United States.	The order of economic liberalization and the informal economy: A review of the literature on the order of economic liberalization and the informal economy.	Can knowledge of financial policy improve the informal economy?	0.051472127
An Inquiry into the Nature and Causes of the Great Earthquake of 1776: A review of the literature on the great earthquake of 1776.	It was the publication in 1776 of the 'An Inquiry into the Nature and Causes of the Great Earthquake of 1776' that led to the development of the informal economy.	Information rules - a strategic guide to the use of information technology: A review of the literature on information rules.	From the Publisher: Information rules is a strategic guide to the use of information technology.	0.346890867
The New Global Economy and the Informal Economy: A review of the literature on the new global economy and the informal economy.	Policy makers in the developing world: This paper discusses the results of a study on the new global economy and the informal economy.	Mental Health: Culture, Race, and the Informal Economy: A review of the literature on mental health, culture, race, and the informal economy.	Mental health is fundamental to the development of the informal economy.	0.088085838
Implicit Racial/Ethnic Bias Among Health Care Providers: A review of the literature on implicit racial/ethnic bias among health care providers.	BACKGROUND In the United States, implicit racial/ethnic bias among health care providers is a major barrier to the development of the informal economy.	Community health workers in the informal economy: A review of the literature on community health workers in the informal economy.	Over the past half-century, community health workers have played a key role in the development of the informal economy.	0.295242786
Integrating Data on Social Determinants of Health: A review of the literature on integrating data on social determinants of health.	As population health becomes a priority, the need for data on social determinants of health is increasing: This paper discusses the results of a study on integrating data on social determinants of health.	A Conceptual and Operational Framework for the Informal Economy: A review of the literature on a conceptual and operational framework for the informal economy.	The acceptance of new information technologies is a key factor in the development of the informal economy.	0.23116371
Information Technology Adoption in the Informal Economy: A review of the literature on information technology adoption in the informal economy.	The process of information technology adoption in the informal economy: This paper discusses the results of a study on information technology adoption in the informal economy.	US Department of Health and Human Services: A review of the literature on the US Department of Health and Human Services.	More than ever before the US Department of Health and Human Services is a key player in the development of the informal economy.	0.042688183
The Impact of COVID-19 on Health: A review of the literature on the impact of COVID-19 on health.	The COVID-19 pandemic has disrupted the informal economy: This paper discusses the results of a study on the impact of COVID-19 on health.	Indoor air pollution in developing countries: A review of the literature on indoor air pollution in developing countries.	Around 50% of people, almost everywhere, are exposed to indoor air pollution.	0.279275328
Location, Competition, and Economic Geography during the Informal Economy: A review of the literature on location, competition, and economic geography during the informal economy.	Economic geography during the informal economy: This paper discusses the results of a study on location, competition, and economic geography during the informal economy.	Time Flies When You're Having Fun: A review of the literature on time flies when you're having fun.	Extant explanations of why users adopt a new technology: This paper discusses the results of a study on time flies when you're having fun.	0.069925487
Applying the Technology Acceptance Model to the Informal Economy: A review of the literature on applying the technology acceptance model to the informal economy.	In this study, we consider the relationship between the technology acceptance model and the informal economy: This paper discusses the results of a study on applying the technology acceptance model to the informal economy.	Understanding and Responding to the Informal Economy: A review of the literature on understanding and responding to the informal economy.	Evidence of a social gradient in health: This paper discusses the results of a study on understanding and responding to the informal economy.	0.03706035
Why cultural safety rather than cultural competence? A review of the literature on why cultural safety rather than cultural competence.	Background Eliminating indigenous health disparities: This paper discusses the results of a study on why cultural safety rather than cultural competence.	National Health Care Spending in the United States: A review of the literature on national health care spending in the United States.	Health care spending in the US is a key factor in the development of the informal economy.	0.193978414
A Resource-Based Perspective on the Informal Economy: A review of the literature on a resource-based perspective on the informal economy.	The resource-based view of the informal economy: This paper discusses the results of a study on a resource-based perspective on the informal economy.	You Say Illegal, I Say Legitimate: A review of the literature on you say illegal, I say legitimate.	The entrepreneurial process drives the development of the informal economy.	0.195501775
Predicting User Intentions: A review of the literature on predicting user intentions.	Information systems IS cannot be used to predict user intentions: This paper discusses the results of a study on predicting user intentions.	Extreme Weather and Climate Change: A review of the literature on extreme weather and climate change.	Extreme weather and climate change are a major threat to the informal economy.	0.000964111
The American Political Economy Part I: Making Sense of the Political Economy: A review of the literature on the American political economy.	Part I: Making Sense of the Political Economy: This paper discusses the results of a study on the American political economy.	Health literacy in Europe: a comparative study: A review of the literature on health literacy in Europe.	Background: Health literacy is a key factor in the development of the informal economy.	0.11124634
Information Technology, Work, and the Informal Economy: A review of the literature on information technology, work, and the informal economy.	Recently, the relative demand for information technology has increased: This paper discusses the results of a study on information technology, work, and the informal economy.	Economic Effects of Coronavirus: A review of the literature on the economic effects of coronavirus.	This report discusses the economic effects of coronavirus.	0.217021257
Stress and the Mental Health of the Informal Economy: A review of the literature on stress and the mental health of the informal economy.	This article provides an overview of the relationship between stress and the mental health of the informal economy: This paper discusses the results of a study on stress and the mental health of the informal economy.	A New Dimension of Health Care: A review of the literature on a new dimension of health care.	Background There is currently a lack of understanding of the relationship between stress and the mental health of the informal economy.	0.176561594
Health literacy as a public health issue: A review of the literature on health literacy as a public health issue.	SUMMARY Health literacy is a key factor in the development of the informal economy: This paper discusses the results of a study on health literacy as a public health issue.	Chemical Pesticides and Human Health: A review of the literature on chemical pesticides and human health.	The industrialization of the agricultural sector has led to the use of chemical pesticides.	0.235305265
Health complaints, stress, and the informal economy: A review of the literature on health complaints, stress, and the informal economy.	Most current models in health psychology do not take into account the relationship between health complaints, stress, and the informal economy: This paper discusses the results of a study on health complaints, stress, and the informal economy.	The influence of education on health: A review of the literature on the influence of education on health.	Background A clear understanding of the influence of education on health is a key factor in the development of the informal economy.	0.249833435
Social Cognitive Theory and the Informal Economy: A review of the literature on social cognitive theory and the informal economy.	A model, based on Bandura's Social Cognitive Theory, is used to understand the relationship between social cognitive theory and the informal economy: This paper discusses the results of a study on social cognitive theory and the informal economy.	The grounded psychometric development of a measure of health literacy: A review of the literature on the grounded psychometric development of a measure of health literacy.	Background Health literacy is a key factor in the development of the informal economy.	0.235751659
Loneliness as a public health issue: A review of the literature on loneliness as a public health issue.	OBJECTIVES We aimed to determine the relationship between loneliness and the informal economy: This paper discusses the results of a study on loneliness as a public health issue.	The role of behavioral science in the development of the informal economy: A review of the literature on the role of behavioral science in the development of the informal economy.	Increasing evidence suggests that the role of behavioral science in the development of the informal economy is a key factor.	0.247409269
Methods for the estimation of the informal economy: A review of the literature on methods for the estimation of the informal economy.	BACKGROUND Cost-effectiveness: This paper discusses the results of a study on methods for the estimation of the informal economy.	Measuring health literacy in the informal economy: A review of the literature on measuring health literacy in the informal economy.	Background Several measurement tools have been used to measure health literacy in the informal economy.	0.365330935
Colonial global economy: towards a new global economy: A review of the literature on the colonial global economy.	Abstract Standard accounts of the colonial global economy: This paper discusses the results of a study on the colonial global economy.	Globalization in Question: The Informal Economy: A review of the literature on globalization in question.	Introduction: globalization - a new global economy: This paper discusses the results of a study on globalization in question.	0.629766345
A conceptual framework for action: A review of the literature on a conceptual framework for action.	When he announced his intention to act, the informal economy was already in existence: This paper discusses the results of a study on a conceptual framework for action.	The outbreak of COVID-19 coronavirus: A review of the literature on the outbreak of COVID-19 coronavirus.	Background: The current outbreak of COVID-19 coronavirus is a major threat to the informal economy.	0.392270446

A	B	C	D	E
The Relationship Between Educ	Adults with higher educational a	Associations of health literacy w	Background Health literacy conc	0.566012144
On the Economy of the Human F	Abstract : An approach to human	Prioritizing the Mental Health ar	The COVID-19 pandemic has hac	0.240585595
Understanding associations amc	Race/ethnicity and socioeconomic	Review: Information Technology	Despite the importance to rese	-0.021087047
The Sustainability of Evidence-B	There is strong interest in imple	Governing the economy : the po	For over one hundred years, the	0.055884778
Mobile medical and health apps	This paper examines the state o	The prescription opioid and heri	Public health authorities have d	0.245327637
Governance in the Blockchain Ec	Blockchain technology is often r	Managerial Ties and Firm Perfor	Using survey data from China, w	0.16661039
Social isolation, loneliness and	The health and well-being conse	The GRADE Evidence to Decision	ObjectiveTo describe a framewi	0.139598802
Health Care Pollution And Publi	An up-to-date assessment of en	Privacy in Context: Technology,	Privacy is one of the most urge	0.08565294
The Political Economy of Terrori	The Political Economy of Terrori	The Role of Telehealth in Reduc	Xiaoyun Zhou, MMH Centaine L	0.021738533
The global burden of oral diseas	This paper outlines the burden c	Differences between caregivers	Providing care for a frail older a	0.16724807
What's Driving the New Econom	Using a unique nationally repres	The World Oral Health Report 20	Chronic diseases and injuries ar	-0.001684289
Valuing health-related quality o	A new version of the EQ-5D, the	Circular Economy and Sustainabi	Textiles are essential to human	0.057044733
Development of an Instrument	This paper reports on the develo	Digital Mental Health and COVID	As interest in and use of telehe	0.216332942
A meta-analysis of self-determi	ABSTRACT There are no literatur	The Mismanagement of Talent: I	The knowledge economy conjun	0.012377407
The SEED and the Rapid Annotat	In 2004, the SEED (http://pubsee	Energy structure, digital econom	As a new production factor, digi	0.007731039
The mental health continuum: fi	This paper introduces and applie	Using Electronic Health Records	The use and functionality of ele	0.19871822
Measuring social class in US pub	Increasing social inequalities in	Knowledge of the Firm, Combin	How should we understand why	-0.002111135
What is the meaning of health li	The objective of this review was	Stigma, mental health, and resil	OBJECTIVES We assessed the as	0.155401319
Blockchain technology and its re	Globalisation of supply chains m	Mental health literacy: empowe	For major physical diseases, it is	0.102674231
Atom Economy-A Challenge for	Enhancing the efficiency of the s	Review of community-based res	Community-based research in p	-0.014982798
Climate change and mental heal	BackgroundThis article provides	The benefits of health informati	An unprecedented federal effo	0.101571828
Principles of Political Economy v	Get Principles Of Political Econo	The rise of the "just-in-time wor	The so-called "gig-economy" ha	0.311520755
Grundrisse : foundations of the	Written during the winter of 185	Using PLS path modeling in new	Purpose - Partial least squares (	0.040736049
Evidence suggesting that a chror	OBJECTIVES This study evaluate	New Tools in Comparative Politi	This article introduces a large ne	0.026315093
A Meta-Analysis of the Effective	The Health Belief Model (HBM; f	The Broad Scope of Health Effic	Background: Concerns for arsen	0.220106214

A	B	C	D	E
Technology Infrastructure and B	The general objective of the stu	Technology Acceptance Model 3	Prior research has provided valu	0.509171128
A framework for public health ar	A 5-tier pyramid best describes	Exposure to Neighborhood Gree	Green space is now widely view	0.350202113
Sleep Health: An Opportunity fo	The concept of sleep health prov	The novel coronavirus (COVID-2	The 2019 novel coronavirus (CO	0.272541702
Evaluating Strategies For Reduci	The opportunities for healthy ch	Implementation of Circular Econ	Small and medium-sized enterp	0.114949159
Monetary Policy and Exchange R	We lay out a small open econom	Effects of soft drink consumption	In a meta-analysis of 88 studies,	0.029463712
The effectiveness of M-health t	BackgroundThe application of m	THE MORAL ECONOMY OF THE EI	WE HAVE BEEN WARNED IN REC	0.029141743
User Acceptance of Information	Abstract Lack of user acceptance	Broad strokes towards a grand th	ABSTRACT This article aims to pr	-0.018175466
The Political Economy of Govern	The determinants of government	Product design and business mo	The transition within business f	-0.002219396
Enhancing treatment fidelity in	Treatment fidelity refers to the	Consumer Acceptance and Use c	This paper extends the unified	0.189796239
Mental illness and/or mental he	A continuous assessment and a	(A Theoretical Extension of the T	The present research develop	0.064266488
Qualitative data analysis for hea	OBJECTIVE To provide practical s	Circular Economy - Challenges f	Abstract The circular economy n	0.117956392
Self-rated health and mortality:	We examine the growing numbe	The National Institute on Minor	We introduce the National Insti	0.401915103
Social and Structural Determinai	Since the World Health Organiza	When patient activation levels c	Patient engagement has become	0.352842897
Compassion and Repression: Thi	Immigration policies in Europe	Consumer health information se	Increasingly, consumers engage	0.046505757
URBAN REGIMES AND THE CAPA	ABSTRACT: Regime theory starts	Discrete choice experiments in	Discrete choice experiments (D	0.174778938
The Structure and Performance	The paper examines the determ	The prevalence, distribution, an	The survey data presented here	0.137830749
The Platform Political Economy	ABSTRACT 'FinTech' is the digital	Patient-centred access to health	BackgroundAccess is central to	0.126248881
Integration of Trade and Disinte	The last few decades have seen	The Role of the Business Model	This paper explores the role of	0.111157149
Mental Health of Children and A	Background: The COVID?19 pand	Bayesian Estimation of an Open	In this paper we develop a dyna	0.047018934
Towards a Circular Economy: The	While there is great potential in	Barriers and facilitators to the us	Limited attention has been paic	0.110024244
Validation and utility of a self-re	CONTEXT The Primary Care Eval	The Globalization of Liberalizati	One of the most important deve	-0.06562563
Meta-analysis of the relationshi	BACKGROUND Risk perceptions	Optimism, coping, and health: a	This article describes a scale me	0.354527593
Translating Social Ecological	The Health promotion programs ofte	Challenges and burden of the Co	Background The coronavirus dis	0.296160281
Farewell to Socialism@@@The	This book provides a comprehen	The Long-Term Labor Market Coi	This paper studies the labor ma	0.231316715
The Effect of Service Employees	The goal of the present paper is	Standards of Care for the Health	Abstract Background: Transgenc	0.193578213

Lampiran 3. Dataset Setelah *Clustering*

A	B	C	D	E	F	G
title1	abstract1	title2	abstract2	pseudo-label	cluster	similarity_level
The Health Utilities Index	This is a review of the He	Perceived Usefulness, Eas	This paper presents the fir	0.180239886	0	Sedikit Berkaitan
Internet paradox. A social	The Internet could chang	Ecological momentary inte	PURPOSE Psychosocial and	0.391628027	3	Cukup Berkaitan
Acculturation and Latino h	This chapter provides an	Dissecting racial bias in an	Racial bias in health algori	0.371246964	3	Cukup Berkaitan
Best Practices for Develop	Scale development and v	'Old Economy' Inputs for 'I	This paper discusses the re	0.054675169	1	Tidak Relevan
Contamination of drinking	The contamination of gro	Linear Economy Versus Cir	Abstract Upon visiting the	0.10084822	1	Tidak Relevan
Collaborative Economy an	House swapping, ridesha	Technology development	A biorefinery that supplen	-0.016242785	1	Tidak Relevan
Toward a Greater Underst	The informal economy co	Evaluating the public heal	Progress in public health a	0.026338331	1	Tidak Relevan
Shaping Agility through Di	Agility is vital to the inno	Police Brutality and Black	We investigated links betw	0.030276757	1	Tidak Relevan
Environment, economy an	Sustainable development	Electronic Data Interchang	Many EDI researchers and	0.139270499	0	Sedikit Berkaitan
Can electronic medical rec	To broadly examine the p	The world economy : histo	Angus Maddison provides	0.077206388	1	Tidak Relevan
The Relationship between	Social cohesion involves	Technology as an occasion	New medical imaging devi	0.134024233	0	Sedikit Berkaitan
The health benefits of urb	BACKGROUND Urban devi	The Duality of Technology	This paper develops a new	0.060369313	1	Tidak Relevan
Perceived Usefulness, Per	Valid measurement scale	Chains and networks, terri	A vast and continually exp	0.001342501	1	Tidak Relevan
What Do We Learn from th	A rapidly growing body of	Hispanic health in the USA	Hispanics are the largest m	0.099247411	1	Tidak Relevan
What the evidence shows	Patient engagement is an	Mental Health Smartphon	Background The number o	0.372995257	3	Cukup Berkaitan
Trasformismo or transform	Abstract What does IPE h	Photovoice: a participator	Photovoice is a participato	0.038940355	1	Tidak Relevan
The Impact of Knowledge-	The aim of this paper is to	Capturing the Complexity	The past decade has broug	0.208763599	0	Sedikit Berkaitan
Socioeconomic disparities	OBJECTIVES We aimed to	Does health literacy medi	While socioeconomic disp	0.753816783	2	Sangat Berkaitan
Rethinking the Informal Ec	This paper explores the re	Antecedents of trust in the	Users and potential users	0.354566872	3	Cukup Berkaitan
Market, Hierarchy, and Tru	Recent conceptualization	The COVID-19 Pandemic a	Abstract The COVID-19 par	0.299443871	3	Cukup Berkaitan
Globalization and history :	Globalization is not a new	Systematic review of the	BackgroundGiven the rapid	-0.008931525	1	Tidak Relevan
User Acceptance of Compu	Computer systems cannot	Mental health problems, c	CONTEXT The US military h	0.044408891	1	Tidak Relevan
Factors Associated With M	Key Points Question Wha	Globalization, the new eco	The globalized new econo	0.061863072	1	Tidak Relevan
Review: Mental health imp	Background: The COVID-1	How Chiefs Come to Powe	1. Introduction: the nature	0.007964596	1	Tidak Relevan
The problem with the phr	Intersectionality is a theo	Health behavior and heal	Health behavior change is	0.435359538	3	Cukup Berkaitan

A	B	C	D	E	F	G
Digital Economy, Technolo	Technological innovation	Mobile text messaging for	The aim of this systematic	0.082428083	1	Tidak Relevan
The Gifts Of Athena Histor	Thank you for reading the	Unequal treatment: confr	Racial and ethnic disparitie	-0.037103318	1	Tidak Relevan
Designing the Business Mc	Switching from the curren	The Global Platform Econ	Global online platforms m	0.342005253	3	Cukup Berkaitan
National Health Care Spen	US health care spending i	The order of economic lib	Can knowledge of financia	0.051472127	1	Tidak Relevan
An Inquiry into the Nature	It was the publication in	Information rules - a strate	From the Publisher: Inform	0.346890867	3	Cukup Berkaitan
The New Global Economy :	Policy makers in the deve	Mental Health: Culture, Re	Mental health is fundamen	0.088085838	1	Tidak Relevan
Implicit Racial/Ethnic Bias	BACKGROUND In the Unit	Community health worker	Over the past half-century	0.295242786	3	Cukup Berkaitan
Integrating Data On Social	As population health bec	A Conceptual and Operati	The acceptance of new inf	0.23116371	0	Sedikit Berkaitan
Information Technology An	The process of informati	US Department of Health	More than ever before the	0.042688183	1	Tidak Relevan
The Impact of COVID-19 or	The COVID-19 pandemic	Indoor air pollution in dev	Around 50% of people, aln	0.279275328	3	Cukup Berkaitan
Location, Competition, an	Economic geography duri	Time Flies When You're H	Extant explanations of wh	0.069925487	1	Tidak Relevan
Applying the Technology A	In this study, we consider	Understanding and Respo	Evidence of a social gradie	0.03706035	1	Tidak Relevan
Why cultural safety rather	BackgroundEliminating in	National Health Care Spen	Health care spending in th	0.193978414	0	Sedikit Berkaitan
A Resource-Based Perspec	The resource-based view	You Say Illegal, I Say Legiti	The entrepreneurial proce	0.195501775	0	Sedikit Berkaitan
Predicting User Intentions	Information systems IS ca	Extreme Weather and Clim	Extreme weather and clim	0.000964111	1	Tidak Relevan
The American Political Eco	Part I: Making Sense of th	Health literacy in Europe:	Background: Health literac	0.11124634	1	Tidak Relevan
Information Technology, V	Recently, the relative der	Economic Effects of Coron	This report discusses the e	0.217021257	0	Sedikit Berkaitan
Stress and the Mental Hea	This article provides an o	A New Dimension of Heal	Background There is curre	0.176561594	0	Sedikit Berkaitan
Health literacy as a public	SUMMARY Health literacy	Chemical Pesticides and H	The industrialization of the	0.235305265	0	Sedikit Berkaitan
Health complaints, stress,	Most current models in h	The influence of education	Background A clear unders	0.249833435	0	Sedikit Berkaitan
Social Cognitive Theory an	A model, based on Bandu	The grounded psychomet	BackgroundHealth literacy	0.235751659	0	Sedikit Berkaitan
Loneliness as a public heal	OBJECTIVES We aimed to	The role of behavioral scie	Increasing evidence sugge	0.247409269	0	Sedikit Berkaitan
Methods for the estimatio	BACKGROUND Cost-effec	Measuring health literacy	BackgroundSeveral measu	0.365330935	3	Cukup Berkaitan
Colonial global economy: I	Abstract Standard accoun	Globalization in Question	Introduction: globalization	0.629766345	2	Sangat Berkaitan
A conceptual framework fo	When he announced his i	The outbreak of COVID-19	Background: The current o	0.392270446	3	Cukup Berkaitan

A	B	C	D	E	F	G
The Relationship Between Adults with higher education	Associations of health literacy	Background Health literacy	0.566012144	2	Sangat Berkaitan	
On the Economy of the Human Abstract : An approach to	Prioritizing the Mental Health	The COVID-19 pandemic health	0.240585595	0	Sedikit Berkaitan	
Understanding association Race/ethnicity and socioeconomic	Review: Information Technology	Despite the importance to	-0.021087047	1	Tidak Relevan	
The Sustainability of Evidence	There is strong interest in	Governing the economy : For over one hundred years	0.055884778	1	Tidak Relevan	
Mobile medical and health	This paper examines the	The prescription opioid and Public health authorities have	0.245327637	0	Sedikit Berkaitan	
Governance in the Blockchain	Blockchain technology is	Managerial Ties and Firm	0.16661039	0	Sedikit Berkaitan	
Social isolation, loneliness	The health and well-being	The GRADE Evidence to Decision	0.139598802	0	Sedikit Berkaitan	
Health Care Pollution And	An up-to-date assessment	Privacy in Context: Technology	0.08565294	1	Tidak Relevan	
The Political Economy of Tele	The Political Economy of	The Role of Telehealth in	0.021738533	1	Tidak Relevan	
The global burden of oral care	This paper outlines the differences between care	Providing care for a frail old	0.16724807	0	Sedikit Berkaitan	
What's Driving the New Economic	Using a unique nationally	The World Oral Health Report	-0.001684289	1	Tidak Relevan	
Valuing health-related quality	A new version of the EQ-5D	Circular Economy and Sustainable	0.057044733	1	Tidak Relevan	
Development of an Instrument	This paper reports on the	Digital Mental Health and	0.216332942	0	Sedikit Berkaitan	
A meta-analysis of self-de	ABSTRACT There are no li	The Mismanagement of The	0.012377407	1	Tidak Relevan	
The SEED and the Rapid An	In 2004, the SEED (http://)	Energy structure, digital e	0.007731039	1	Tidak Relevan	
The mental health continu	This paper introduces an	Using Electronic Health Re	0.19871822	0	Sedikit Berkaitan	
Measuring social class in U	Increasing social inequality	Knowledge of the Firm, C	-0.002111135	1	Tidak Relevan	
What is the meaning of he	The objective of this review	Stigma, mental health, an	0.155401319	0	Sedikit Berkaitan	
Blockchain technology and	Globalisation of supply ch	Mental health literacy: en	0.102674231	1	Tidak Relevan	
Atom Economy-A Challenge	Enhancing the efficiency	Review of community-based	-0.014982798	1	Tidak Relevan	
Climate change and mental	BackgroundThis article pr	The benefits of health inf	0.101571828	1	Tidak Relevan	
Principles of Political Econ	Get Principles Of Political	The rise of the "just-in-ti	0.311520755	3	Cukup Berkaitan	
Grundrisse : foundations o	Written during the winter	Using PLS path modeling i	0.040736049	1	Tidak Relevan	
Evidence suggesting that a	OBJECTIVES This study ev	New Tools in Comparative	0.026315093	1	Tidak Relevan	
A Meta-Analysis of the Eff	The Health Belief Model	The Broad Scope of Health	0.220106214	0	Sedikit Berkaitan	

A	B	C	D	E	F	G
Technology Infrastructure	The general objective of	Technology Acceptance M	Prior research has provide	0.509171128	2	Sangat Berkaitan
A framework for public he	A 5-tier pyramid best des	Exposure to Neighborhoo	Green space is now widely	0.350202113	3	Cukup Berkaitan
Sleep Health: An Opportu	The concept of sleep heal	The novel coronavirus (CC	The 2019 novel coronavirus	0.272541702	0	Sedikit Berkaitan
Evaluating Strategies For	The opportunities for hea	Implementation of Circu	Small and medium-sized e	0.114949159	1	Tidak Relevan
Monetary Policy and Excha	We lay out a small open e	Effects of soft drink consu	In a meta-analysis of 88 stu	0.029463712	1	Tidak Relevan
The effectiveness of M-he	The applicatio	THE MORAL ECONOMY OF	WE HAVE BEEN WARNED IN	0.029141743	1	Tidak Relevan
User Acceptance of Inform	Abstract Lack of user acce	Broad strokes towards a g	ABSTRACT This article aims	-0.018175466	1	Tidak Relevan
The Political Economy of G	The determinants of gove	Product design and busine	The transition within busin	-0.002219396	1	Tidak Relevan
Enhancing treatment fidel	Treatment fidelity refers	Consumer Acceptance and	This paper extends the uni	0.189796239	0	Sedikit Berkaitan
Mental illness and/or men	A continuous assessment	A Theoretical Extension of	The present research deve	0.064266488	1	Tidak Relevan
Qualitative data analysis f	OBJECTIVE To provide pra	Circular Economy - Challer	Abstract The circular econc	0.117956392	1	Tidak Relevan
Self-rated health and mor	We examine the growing	The National Institute on	We introduce the National	0.401915103	3	Cukup Berkaitan
Social and Structural Deter	Since the World Health O	When patient activation le	Patient engagement has b	0.352842897	3	Cukup Berkaitan
Compassion and Repressic	Immigration policies in E	Consumer health informat	Increasingly, consumers e	0.046505757	1	Tidak Relevan
URBAN REGIMES AND THE	ABSTRACT: Regime theor	Discrete choice experime	Discrete choice experime	0.174778938	0	Sedikit Berkaitan
The Structure and Perform	The paper examines the c	The prevalence, distributi	The survey data presented	0.137830749	0	Sedikit Berkaitan
The Platform Political Ecor	ABSTRACT 'FinTech' is the	Patient-centred access to	BackgroundAccess is centr	0.126248881	0	Sedikit Berkaitan
Integration of Trade and D	The last few decades have	The Role of the Business	This paper explores the ro	0.111157149	1	Tidak Relevan
Mental Health of Children	Background: The COVID-19	Bayesian Estimation of an	In this paper we develop a	0.047018934	1	Tidak Relevan
Towards a Circular Econom	While there is great pote	Barriers and facilitators to	Limited attention has been	0.110024244	1	Tidak Relevan
Validation and utility of a	CONTEXT The Primary Car	The Globalization of Liber	One of the most important	-0.06562563	1	Tidak Relevan
Meta-analysis of the relati	BACKGROUND Risk perce	Optimism, coping, and he	This article describes a sca	0.354527593	3	Cukup Berkaitan
Translating Social Ecologic	Health promotion progr	Challenges and burden of	Background The coronavir	0.296160281	3	Cukup Berkaitan
Farewell to Socialism@@@	This book provides a com	The Long-Term Labor Mark	This paper studies the labo	0.231316715	0	Sedikit Berkaitan
The Effect of Service Empli	The goal of the present p	Standards of Care for the	Abstract Background: Tran	0.193578213	0	Sedikit Berkaitan

Lampiran 4. Source Code

```
from fastapi import FastAPI
from pydantic import BaseModel
from sentence_transformers import SentenceTransformer,
util
import joblib, json, re, numpy as np

app = FastAPI()
model = SentenceTransformer("all-mpnet-base-v2")

kmeans = joblib.load("./model/kmeans_model.joblib")

with open("./model/cluster_label_mapping.json", "r") as f:
    raw_map = json.load(f)
    cluster_label_mapping = {int(k): v for k, v in
raw_map.items()}

bilingual_labels = {
    "Tidak Relevan": "Not Relevant",
    "Sedikit Berkaitan": "Slightly Related",
    "Cukup Berkaitan": "Moderately Related",
    "Sangat Berkaitan": "Highly Related"
}

class TextPair(BaseModel):
    text1: str
    text2: str

def preprocess_text(text: str) -> str:
    text = re.sub(r"\n", " ", text)
    text = re.sub(r"\s+", " ", text)
    return text.strip()

@app.post("/similarity/")
async def calculate_similarity(data: TextPair):
    t1 = preprocess_text(data.text1)
    t2 = preprocess_text(data.text2)

    print(f"Processing texts:\nText 1: {t1}\nText 2:
{t2}")
```

```

emb1 = model.encode(t1, convert_to_tensor=True,
normalize_embeddings=True)
emb2 = model.encode(t2, convert_to_tensor=True,
normalize_embeddings=True)
sim = util.pytorch_cos_sim(emb1, emb2).item()

cluster_idx = int(kmeans.predict([[sim]])[0])
label_kemiripan =
cluster_label_mapping[cluster_idx]
label_english =
bilingual_labels.get(label_kemiripan, label_kemiripan)

return {
    "similarity_score": sim,
    "cluster": cluster_idx,
    "label_kemiripan": label_kemiripan,
    "label_english": label_english,
    "bilingual_labels": {
        "indonesian": {
            "scoreLabel": "Skor Kemiripan",
            "similarity": label_kemiripan
        },
        "english": {
            "scoreLabel": "Similarity Score",
            "similarity": label_english
        }
    }
}

}

=== PSEUDO-LABELING ===
import pandas as pd
from sentence_transformers import SentenceTransformer,
util
from tqdm.auto import tqdm
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy import stats

model = SentenceTransformer("all-mpnet-base-v2")

df = pd.read_csv("./dataset/dataset-journal.csv",
encoding="latin-1")

```

```

all_sentences = pd.concat([df["abstract1"],
df["abstract2"]]).unique().tolist()
sentence_to_emb = {}
for sent, emb in zip(all_sentences,
model.encode(all_sentences, show_progress_bar=True,
convert_to_tensor=True)):
    sentence_to_emb[sent] = emb

pseudo_scores = []
for a, b in tqdm(zip(df["abstract1"], df["abstract2"]),
total=len(df)):
    emb_a = sentence_to_emb[a]
    emb_b = sentence_to_emb[b]
    score = util.cos_sim(emb_a, emb_b).item()

    pseudo_scores.append(score)

df["pseudo-label"] = pseudo_scores

=== Evaluasi Clustering dengan KMeans ===
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score,
davies_bouldin_score

df = pd.read_csv('./dataset/journal-pseudo-label.csv',
encoding='latin-1')
X = df[['pseudo-label']].values

inertia = []
silhouette = []
db_scores = []
k_range = range(2, 9)

print("Evaluasi Clustering untuk berbagai jumlah
cluster:")
print("-" * 60)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    labels = kmeans.labels_

```

```

inertia_val = kmeans.inertia_
silhouette_val = silhouette_score(X, labels)
db_score = davies_bouldin_score(X, labels)

inertia.append(inertia_val)
silhouette.append(silhouette_val)
db_scores.append(db_score)

print(f"k={k:2d} | Inertia: {inertia_val:8.2f} |
Silhouette: {silhouette_val:.4f} | DB Score:
{db_score:.4f}")

=== Clustering Pseudo-label ===
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
import json

import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
import json

# # Load data
df = pd.read_csv('./dataset/journal-pseudo-label.csv',
encoding='latin-1')
X = df[['pseudo-label']].values

# Hasil evaluasi menunjukkan k=4 optimal berdasarkan 3
metrik evaluasi
optimal_clusters = 4
print(f"Melakukan clustering dengan {optimal_clusters}
cluster")
print("    (Jumlah optimal berdasarkan evaluasi 3
metrik)")

# Clustering final

```

```

final_kmeans = KMeans(n_clusters=optimal_clusters,
random_state=42).fit(X)
df['cluster'] = final_kmeans.labels_

print(f"Clustering selesai dengan {optimal_clusters}
cluster")

# Urutkan centroid untuk mapping label
centers = final_kmeans.cluster_centers_.flatten()
sorted_idx = np.argsort(centers)

# Mapping ke label sesuai urutan centroid (rendah ke
tinggi)
labels = ["Tidak Relevan", "Sedikit Berkaitan", "Cukup
Berkaitan", "Sangat Berkaitan"]
cluster_map = {
    int(sorted_idx[i]): labels[i]
    for i in range(len(labels))
}

# Terapkan mapping label ke data
df['similarity_level'] = df['cluster'].map(cluster_map)

# Simpan hasil clustering
df.to_csv('./dataset/journal-clustered.csv',
index=False)
print("Hasil clustering disimpan di './dataset/journal-
clustered.csv'")

# Simpan model clustering
joblib.dump(final_kmeans,
'./model/kmeans_model.joblib')
print("Model clustering disimpan di
'./model/kmeans_model.joblib'")

=== Integrasi Google Gemini API untuk menghasilkan
respons kreatif ===
import { GoogleGenAI } from "@google/genai";

const ai = new GoogleGenAI({
    apiKey: process.env.GEMINI_API_KEY,
});

export async function generateCreativeResponseGemini(

```

```

text1,
text2,
similarityScore,
labelKemiripan
) {
  const system =
    "You are an AI for journal comparison. Analyze the
two journal abstracts and provide a comprehensive
comparison in both Indonesian and English.";

  const content = `
    Similarity Label: ${labelKemiripan}
    Similarity Score: ${similarityScore}
    Abstract 1: ${text1}
    Abstract 2: ${text2}

    Analyze and compare these journals.`;

  // Define the response schema
  const responseSchema = {
    type: "object",
    properties: {
      indonesian: {
        type: "object",
        properties: {
          title: { type: "string" },
          firstJournalSummary: { type: "string" },
          secondJournalSummary: { type: "string" },
          similarities: {
            type: "array",
            items: { type: "string" },
          },
          differences: {
            type: "array",
            items: { type: "string" },
          },
          relationshipAnalysis: { type: "string" },
          conclusion: { type: "string" },
        },
      },
      required: [
        "title",
        "firstJournalSummary",
        "secondJournalSummary",
        "similarities",
        "differences",

```



```

        "relationshipAnalysis",
        "conclusion",
    ],
},
english: {
    type: "object",
    properties: {
        title: { type: "string" },
        firstJournalSummary: { type: "string" },
        secondJournalSummary: { type: "string" },
        similarities: {
            type: "array",
            items: { type: "string" },
        },
        differences: {
            type: "array",
            items: { type: "string" },
        },
        relationshipAnalysis: { type: "string" },
        conclusion: { type: "string" },
    },
    required: [
        "title",
        "firstJournalSummary",
        "secondJournalSummary",
        "similarities",
        "differences",
        "relationshipAnalysis",
        "conclusion",
    ],
},
required: ["indonesian", "english"],
};

try {
    const response = await ai.models.generateContent({
        model: "gemini-2.0-flash",
        contents: content,
        config: {
            systemInstruction: system,
            temperature: 0.7,
            maxOutputTokens: 1500,
            responseMimeType: "application/json",
            responseSchema: responseSchema,

```

```

    },
  });

  const responseText = response.text;
  const parsedResponse = JSON.parse(responseText);

  // Format the response into markdown
  const formattedResponse = {
    indonesian:
formatToMarkdown(parsedResponse.indonesian,
"indonesian"),
    english: formatToMarkdown(parsedResponse.english,
"english"),
  };

  return formattedResponse;
} catch (error) {
  console.error("Error calling Gemini:", error);
  throw new Error("Failed to generate creative
response");
}
}

function formatToMarkdown(data, language) {
  const translations = {
    indonesian: {
      firstJournalTitle: "## 1. Ringkasan Jurnal
Pertama",
      secondJournalTitle: "## 2. Ringkasan Jurnal
Kedua",
      similaritiesTitle: "## Kesamaan Utama:",
      differencesTitle: "## Perbedaan Utama:",
      analysisTitle: "## Analisis Hubungan",
      conclusionTitle: "## Kesimpulan",
    },
    english: {
      firstJournalTitle: "## 1. First Journal Summary",
      secondJournalTitle: "## 2. Second Journal
Summary",
      similaritiesTitle: "## Main Similarities:",
      differencesTitle: "## Main Differences:",
      analysisTitle: "## Relationship Analysis",
      conclusionTitle: "## Conclusion",
    },
  };
};

```

```

const t = translations[language];

let markdown = `# ${data.title}\n\n`;
markdown +=
`${t.firstJournalTitle}\n${data.firstJournalSummary}\n\n`;
markdown +=
`${t.secondJournalTitle}\n${data.secondJournalSummary}\n\n`;

markdown += `${t.similaritiesTitle}\n`;
data.similarities.forEach((similarity) => {
  markdown += `- ${similarity}\n`;
});
markdown += `\n`;

markdown += `${t.differencesTitle}\n`;
data.differences.forEach((difference) => {
  markdown += `- ${difference}\n`;
});
markdown += `\n`;

markdown +=
`${t.analysisTitle}\n${data.relationshipAnalysis}\n\n`;
markdown +=
`${t.conclusionTitle}\n${data.conclusion}`;

return markdown;
}

```

Lampiran 5. Surat Keterangan Bebas Plagiat



**MAJELIS PENDIDIKAN TINGGI PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH MAKASSAR
UPT PERPUSTAKAAN DAN PENERBITAN**
Alamat kantor: Jl. Sultan Alauddin NO.259 Makassar 90221 Tlp.(0411) 866972,881593, Fax.(0411) 865588

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

SURAT KETERANGAN BEBAS PLAGIAT

**UPT Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar,
Menerangkan bahwa mahasiswa yang tersebut namanya di bawah ini:**

Nama : Muhammad Asygar Faeruddin
Nim : 105841104721
Program Studi : Teknik Informatika

Dengan nilai:

No	Bab	Nilai	Ambang Batas
1	Bab 1	9%	10 %
2	Bab 2	2%	25 %
3	Bab 3	8%	10 %
4	Bab 4	0%	10 %
5	Bab 5	5%	5 %

Dinyatakan telah lulus cek plagiat yang diadakan oleh UPT- Perpustakaan dan Penerbitan Universitas Muhammadiyah Makassar Menggunakan Aplikasi Turnitin.

Demikian surat keterangan ini diberikan kepada yang bersangkutan untuk dipergunakan seperlunya.

Makassar, 26 Agustus 2025
Mengetahui,
Kepala UPT- Perpustakaan dan Penerbitan,



Nursinah, S.Hum., M.I.P
NBM. 964 591

Jl. Sultan Alauddin no 259 makassar 90222
Telepon (0411)866972,881 593,fax (0411)865 588
Website: www.library.unismuh.ac.id
E-mail : perpustakaan@unismuh.ac.id

Lampiran 6. Uji Plagiat BAB 1-5

BAB I Muhammad Asygar
Faeruddin 105841104721
by Tahap Tutup



Submission date: 25-Aug-2025 11:00AM (UTC+0700)

Submission ID: 2734752188

File name: BAB_1_5.docx (44.6K)

Word count: 949

Character count: 6711

BAB I Muhammad Asygar Faeruddin 105841104721

ORIGINALITY REPORT

9%	5%	2%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Muhammadiyah Makassar Student Paper	5%
2	repositori.unsil.ac.id Internet Source	3%
3	Arief Rahman Hakim, Alva Hendi Muhammad. "PERBANDINGAN MODEL TRANSFORMER, DEEP LEARNING, DAN MACHINE LEARNING UNTUK DETEKSI BERITA PALSU: STUDI KASUS PADA TEKS BERBAHASA INDONESIA", Jurnal Manajemen Informatika dan Sistem Informasi, 2025 Publication	2%

Exclude quotes On
Exclude bibliography On

Exclude matches < 2%

BAB II Muhammad Asygar Faeruddin 105841104721

by Tahap Tutup

Submission date: 25-Aug-2025 11:02AM (UTC+0700)
Submission ID: 2734754042
File name: BAB_2_3.docx (121.75K)
Word count: 2044
Character count: 13996

BAB II Muhammad Asygar Faeruddin 105841104721

ORIGINALITY REPORT

2%

SIMILARITY INDEX

2%

INTERNET SOURCES

0%

PUBLICATIONS

0%

STUDENT PAPERS

PRIMARY SOURCES

1

digilibadmin.unismuh.ac.id

Internet Source

2%

Exclude quotes

On

Exclude matches

< 2%

Exclude bibliography

On



BAB III Muhammad Asygar Faeruddin 105841104721

by Tahap Tutup



Submission date: 25-Aug-2025 11:03AM (UTC+0700)

Submission ID: 2734754736

File name: BAB_3_4.docx (108.08K)

Word count: 1040

Character count: 7155

BAB III Muhammad Asygar Faeruddin 105841104721

ORIGINALITY REPORT

8%

SIMILARITY INDEX

7%

INTERNET SOURCES

0%

PUBLICATIONS

7%

STUDENT PAPERS

PRIMARY SOURCES

1

dac.telkomuniversity.ac.id

Internet Source

5%

2

Submitted to Universitas Muhammadiyah
Makassar

Student Paper

3%

Exclude quotes

On

Exclude bibliography

On

Exclude matches

2%

BAB IV Muhammad Asygar Faeruddin 105841104721

by Tahap Tutup



Submission date: 25-Aug-2025 11:05AM (UTC+0700)

Submission ID: 2734756705

File name: BAB_4_2.docx (1.62M)

Word count: 2698

Character count: 17556

BAB IV Muhammad Asygar Faeruddin 105841104721

ORIGINALITY REPORT

0%

SIMILARITY INDEX

0%

INTERNET SOURCES

0%

PUBLICATIONS

0%

STUDENT PAPERS

PRIMARY SOURCES

Exclude quotes On

Exclude bibliography On

Exclude matches < 2%



BAB V Muhammad Asygar Faeruddin105841104721

by Tahap Tutup



Submission date: 25-Aug-2025 11:06AM (UTC+0700)

Submission ID: 2734757413

File name: BAB_5_2.docx (31.56K)

Word count: 313

Character count: 2266

BAB V Muhammad Asygar Faeruddin105841104721

ORIGINALITY REPORT

5%

SIMILARITY INDEX

5%

INTERNET SOURCES

5%

PUBLICATIONS

0%

STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

3%

★ pub.unj.ac.id

Internet Source

Exclude quotes

On

Exclude bibliography

On

Exclude matches

< 2%

