Cerdika: Jurnal Ilmiah Indonesia, 5 (12), 2025
p-ISSN: 2774-6291 e-ISSN: 2774-6534
Available online at http://cerdika.publikasiindonesia.id/index.php/cerdika/index

Implementasi Model Sentence-Bert untuk Deteksi Plagiarisme Pada Karya Tulis Ilmiah Psikologi Berbahasa Indonesia

Ilhan Hakiki, Firas Atqiya, Ahmad Suryan
Universitas Muhammadiyyah Bandung, Indonesia
Email: ilhan.hakiki@umbandung.ac.id, firasatqiya@umbandung.ac.id, suryan@umbandung.ac.id

Abstrak
Plagiarisme merupakan permasalahan serius dalam dunia akademik, khususnya pada karya tulis ilmiah berbahasa Indonesia di bidang psikologi. Bidang psikologi merupakan salah satu bidang yang rentan terhadap plagiarisme. Penelitian ini bertujuan untuk mengimplementasikan metode Sentence-BERT dalam mendeteksi tingkat kemiripan antar teks guna mengidentifikasi potensi plagiarisme. Dataset yang digunakan terdiri dari 15 pasang abstrak karya tulis ilmiah psikologi yang dikategorikan ke dalam lima jenis, yaitu copy-paste murni, parafrase kuat, mosaik, topik berbeda, dan topik mirip. Setiap pasangan data dibandingkan menggunakan Sentence-BERT untuk menghasilkan embedding semantik, lalu dihitung tingkat kemiripannya dengan cosine similarity. Hasil pengujian menunjukkan bahwa metode Sentence-BERT cukup efektif mendeteksi plagiarisme dengan akurasi global 53,3%, terutama pada kasus copy-paste murni (100%). Namun, kelemahan ditemukan pada kategori mosaik (0%) dan topik mirip (33,3%), di mana model masih kesulitan membedakan teks yang memiliki kesamaan semantik meski berbeda tujuan atau struktur. Dengan demikian, Sentence-BERT terbukti unggul dalam mendeteksi plagiarisme eksplisit, tetapi perlu pengembangan lebih lanjut agar mampu mengenali plagiarisme kompleks.
Kata kunci: Sentence-BERT, Cosine Similarity, Plagiarisme, Psikologi, Deteksi Teks

Abstract
Plagiarism is a serious problem in the academic world, especially in Indonesian-language scientific papers in the field of psychology. The field of psychology is one of the disciplines that is vulnerable to plagiarism. This study aims to implement the Sentence-BERT method in detecting the level of similarity between texts to identify potential plagiarism. The dataset used consists of 15 pairs of psychology abstracts categorized into five types: pure copy-paste, strong paraphrase, mosaic, different topics, and similar topics. Each data pair was compared using Sentence-BERT to generate semantic embeddings, and then the similarity level was calculated using cosine similarity. The test results show that the Sentence-BERT method is quite effective in detecting plagiarism with a global accuracy of 53.3%, especially in pure copy-paste cases (100%). However, weaknesses were found in the mosaic category (0%) and similar topics (33.3%), where the model still struggles to distinguish texts with high semantic similarity even though they have different purposes or structures. Thus, Sentence-BERT proves to be effective in detecting explicit plagiarism, but further development is needed to be able to recognize more complex plagiarism cases.
Keywords: *Sentence-BERT, Cosine Similarity, Plagiarism, Psychology, Text Detection*

PENDAHULUAN

Plagiarisme telah menjadi ancaman serius terhadap integritas akademik global, terutama di era digital yang memfasilitasi akses dan duplikasi informasi dengan sangat mudah (Horbach & Halffman, 2019; Pratiwi & Aisya, 2021). Di era modern ini perkembangan teknologi informasi terutama teknologi digital sudah banyak dimanfaatkan oleh masyarakat dan sudah menjadi salah satu kebutuhan yang tidak dapat dipisahkan dari kehidupan manusia modern (Azmi, 2022). Teknologi digital sudah banyak digunakan dalam bidang ekonomi (Susilo et al., 2021), kesehatan (Purba et al., 2024), industri (Wulansari et al., 2022), pendidikan (Munawar et al., 2021), dan bidang-bidang lainnya.

Dalam konteks pendidikan tinggi Indonesia, transformasi dokumen akademik ke format digital telah menciptakan paradoks: di satu sisi meningkatkan aksesibilitas informasi ilmiah, namun di sisi lain membuka celah lebih luas bagi praktik plagiarisme (Shadiqi, 2019). Salah

satu implementasi teknologi digital dalam bidang pendidikan adalah dokumen digital (Ambarwati et al., 2021; Fad'li et al., 2023). Saat ini banyak sekali dokumen yang sudah diubah dalam bentuk digital, hal ini memudahkan masyarakat dalam menyimpan dokumen yang banyak, memudahkan dalam pencarian dokumen, lebih efektif dan efisien untuk dibawa dan digunakan kapan saja dan di mana saja.

Dokumen digital memang memiliki banyak kelebihan, namun tidak dapat dipungkiri bahwa dokumen dalam bentuk digital ini juga membuat tantangan baru bagi masyarakat terutama mahasiswa. Dokumen dalam bentuk digital sangat mudah diakses dan diduplikasi atau dijiplak. Penjiplakan ini lebih sering dikenal dengan sebutan plagiarisme. Menurut Indra Charismiadji selaku Pengamat dan Praktisi Pendidikan mengatakan bahwa di Indonesia tindakan plagiarisme, khususnya pada karya tulis sudah bukan termasuk hal yang langka. Bahkan tindakan plagiarisme sudah dimulai dari jenjang sekolah dasar hingga menengah dengan angka mencapai 94% (Salbiah, 2021). Fenomena ini mengindikasikan bahwa plagiarisme bukan hanya masalah teknis deteksi, tetapi juga mencerminkan krisis pemahaman etika akademik yang perlu diatasi melalui pendekatan teknologi dan edukasi terintegrasi.

Praktik penjiplakan atau plagiarisme ini sangat mudah dilakukan dengan teknik copy-paste-modify sebagian atau seluruh isi dokumen. Di kalangan perguruan tinggi, praktik plagiarisme sering terjadi karena mahasiswa yang sudah terbiasa mengambil tulisan dari karya orang lain tanpa mencantumkan sitasi atau sumbernya. Praktik plagiarisme ini sering terjadi ketika mahasiswa sedang mengerjakan tugas-tugas biasa, tugas akhir atau skripsi (Azmi, 2022).

Ada dua faktor yang membuat masyarakat melakukan tindakan plagiarisme, yaitu kurangnya pemahaman penulis dalam metode penulisan karya tulis ilmiah dan rasa tidak pedulinya penulis dalam menerapkan kode etik penulisan karya tulis ilmiah (Sinurat et al., 2021). Menurut Bruton dalam (Sinurat et al., 2021) mengatakan bahwa plagiarisme ini adalah tindakan yang tidak sengaja dan akan mendapatkan hukuman jika plagiarisme yang dilakukan mempunyai dampak yang besar. Menurut (Hanum et al., 2021), terdapat banyak faktor yang menyebabkan penulis melakukan pelanggaran tersebut, seperti kurangnya pemahaman tentang etika pengutipan, cara menyitasi, kelalaian dalam mencantumkan sumber referensi yang telah digunakan, dan sebagainya.

Pelaku yang terbukti telah melakukan tindakan plagiarisme harus siap menghadapi konsekuensi hukum, termasuk pencabutan gelar sesuai UU No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional Pasal 25 Ayat 2, pembatalan ijazah berdasarkan Peraturan Menteri Pendidikan dan Kebudayaan No. 17 Tahun 2010 Pasal 12 Ayat 1 Huruf g, serta ancaman hukuman pidana penjara maksimal dua tahun dan/atau denda hingga Rp 200.000.000 (dua ratus juta rupiah) sesuai UU No. 20 Tahun 2003 Pasal 70 (Hanum et al., 2021).

Plagiarisme memiliki dampak negatif yang luas, baik bagi individu maupun institusi. Bagi mahasiswa atau peneliti, plagiarisme dapat menyebabkan sanksi akademis, seperti penurunan nilai atau bahkan dikeluarkan dari institusi. Bagi institusi pendidikan, kasus plagiarisme dapat merusak reputasi dan kredibilitas. Di lingkungan profesional, tindakan plagiarisme dapat merusak citra individu maupun organisasi. Masalah ini pernah menjadi isu serius di kalangan ilmuwan Tiongkok. Sebuah studi mengungkapkan bahwa plagiarisme tetap menjadi tantangan meskipun telah ada berbagai upaya untuk mengatasinya. National Science Foundation of China (NSFC) melaporkan bahwa 34% ilmuwan yang menerima dana pemerintah terbukti melakukan plagiarisme antara tahun 1999 hingga 2005 (Hestiani &

Suriyani, 2023). Selain itu, plagiarisme juga menghambat perkembangan ilmu pengetahuan karena mengurangi orisinalitas dan inovasi. Oleh karena itu, deteksi plagiarisme yang cepat dan akurat menjadi salah satu kebutuhan, terutama di era digital yang penuh dengan informasi mudah diakses.

Bidang psikologi merupakan salah satu disiplin ilmu yang sangat rentan terhadap plagiarisme karena karakteristik penulisannya yang sering melibatkan deskripsi konsep teoritis, tinjauan literatur ekstensif, dan pembahasan fenomena psikologis yang cenderung menggunakan terminologi dan kerangka berpikir yang serupa (Ekaningtyas, 2022; Sugiarto et al., 2021). Plagiarisme menjadi masalah yang serius di kalangan mahasiswa saat ini karena dapat merusak nilai-nilai kejujuran dan orisinalitas dalam dunia pendidikan dan profesional. Selain karena kurangnya pemahaman mengenai etika penulisan, plagiarisme juga dapat disebabkan karena keterbatasan alat untuk mendeteksi plagiarisme yang akurat dan mudah untuk di akses, terutama bagi institusi pendidikan dengan sumber daya terbatas. Alat deteksi plagiarisme yang ada saat ini, seperti Turnitin atau Grammarly, sering kali memerlukan biaya langganan yang mahal dan tidak selalu terjangkau oleh semua pihak. Selain itu, beberapa alat tersebut kurang efektif dalam mendeteksi plagiarisme pada dokumen berbahasa Indonesia atau dokumen dengan struktur kalimat yang kompleks.

Penelitian terdahulu telah mengeksplorasi berbagai pendekatan untuk deteksi plagiarisme, namun masih terdapat gap signifikan yang perlu diisi. Beberapa penelitian telah dilakukan untuk mengembangkan sistem deteksi plagiarisme. Pawestri & Suyanto (2024) melakukan studi komparatif antara tiga metode similarity: Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Hasil penelitian menunjukkan bahwa cosine similarity mengungguli kedua metode lainnya dengan rata-rata performa lebih baik dalam menentukan plagiarisme dokumen. Namun, penelitian tersebut memiliki keterbatasan karena hanya menguji pada level dokumen penuh tanpa analisis granular per kalimat, sehingga tidak mampu mengidentifikasi bagian spesifik yang mengalami plagiarisme. Misalnya pada penelitian (Pawestri & Suyanto, 2024) mencoba membandingkan tiga metode similarity untuk mendeteksi kemiripan dokumen, yaitu Jaccard Coefficient, Cosine Similarity, dan Euclidean Distance. Dari penelitian tersebut disimpulkan bahwa cosine similarity lebih baik dalam menentukan plagiarisme dalam dokumen karena memiliki rata-rata performa yang lebih baik daripada Jaccard Coefficient dan Euclidean Distance.

Dalam penelitian (Herlambang et al., 2021) mengembangkan sistem deteksi plagiarisme berbasis web dengan metode cosine similarity. Namun, sistem tersebut masih memiliki keterbatasan dalam hal ketika memberikan persentasi kemiripan dokumen tidak disertai dengan detail informasi pada bagian yang memiliki kemiripan. Metode Sentence-BERT dapat menghasilkan embedding semantik yang dapat dibandingkan per kalimat sehingga dapat memberikan informasi terkait kalimat manakah yang memiliki kemiripan.

Untuk mengatasi gap tersebut, penelitian ini mengadopsi pendekatan berbasis deep learning melalui Sentence-BERT (Devlin et al., 2019), sebuah varian BERT yang dioptimalkan khusus untuk menghasilkan sentence embeddings berkualitas tinggi. Berbeda dengan TF-IDF atau metode berbasis kata tradisional, Sentence-BERT mampu menangkap representasi semantik kontekstual yang lebih kaya, memungkinkan deteksi kemiripan makna meskipun terjadi variasi sintaksis atau parafrase (Reimers & Gurevych, 2019). Lebih penting lagi, Sentence-BERT memungkinkan komparasi tingkat kalimat (sentence-level comparison),

sehingga dapat mengidentifikasi secara presisi bagian-bagian teks yang mengalami plagiarisme—sebuah fitur yang absen dalam penelitian-penelitian sebelumnya.

Kebaruan (novelty) penelitian ini terletak pada tiga aspek utama: (1) Penggunaan Sentence-BERT untuk deteksi plagiarisme pada dokumen akademik berbahasa Indonesia di bidang psikologi, yang sebelumnya belum banyak dieksplorasi; (2) Pengembangan taksonomi plagiarisme yang komprehensif mencakup lima kategori (copy-paste murni, parafrase kuat, mosaik, topik berbeda, dan topik mirip) yang memungkinkan evaluasi performa model pada berbagai tingkat kompleksitas plagiarisme; dan (3) Integrasi sistem deteksi level kalimat yang memungkinkan identifikasi presisi bagian-bagian spesifik yang terindikasi plagiarisme, melampaui keterbatasan sistem berbasis skor keseluruhan dokumen dalam penelitian terdahulu.

Penelitian ini bertujuan untuk mengimplementasikan metode sentence-BERT pada alat deteksi kemiripan antar teks. Hasil dari penelitian ini diharapkan dapat menjadi solusi yang efektif dan dapat diintegrasikan ke dalam berbagai sistem, khususnya untuk kebutuhan akademik di Universitas Muhammadiyah Bandung. Dengan menggunakan metode sentence-BERT, sistem ini diharapkan dapat mendeteksi kemiripan antar teks secara akurat, bahkan untuk dokumen berbahasa Indonesia. Selain itu, model ini akan dievaluasi berdasarkan performa dalam mendeteksi plagiarisme, seperti tingkat akurasi dan efisiensi dalam membandingkan teks. Dengan demikian, penelitian ini diharapkan dapat berkontribusi dalam meningkatkan integritas akademis serta memberikan dasar bagi pengembangan sistem deteksi plagiarisme yang lebih canggih di masa depan.

METODE PENELITIAN

Analisis kebutuhan sistem dilakukan untuk menentukan perangkat keras dan perangkat lunak yang diperlukan dalam membangun sistem deteksi plagiarisme. Untuk perangkat keras, digunakan laptop atau PC dengan spesifikasi AMD Ryzen 5 5500U, RAM 8GB, penyimpanan SSD 512 GB, serta koneksi internet yang diperlukan untuk mengunduh dataset, library, dan tools pendukung. Sedangkan untuk perangkat lunak, sistem operasi yang digunakan adalah Windows 11, dengan lingkungan pengembangan Google Colaboratory (Colab) sebagai platform berbasis cloud. Bahasa pemrograman yang digunakan adalah Python versi 3.8 atau lebih baru, dengan beberapa library seperti Pandas untuk mengolah dataset CSV, Numpy untuk operasi matematika, Scikit-learn untuk perhitungan cosine similarity dan evaluasi model, serta sentence-transformers untuk embedding teks. Framework Streamlit juga digunakan untuk pembuatan tampilan. Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk mengimplementasikan model deteksi plagiarisme menggunakan metode Sentence-BERT, berfokus pada abstrak dokumen akademik berbahasa Indonesia di bidang psikologi. Model ini akan diuji dengan 25 pasang sampel dokumen yang dibandingkan berdasarkan abstrak karya tulis ilmiah. Metodologi penelitian terdiri dari beberapa tahap utama, termasuk studi literatur, pengumpulan dataset dan ground truth, preprocessing data, representasi vektor dengan Sentence-BERT, perhitungan cosine similarity, evaluasi model dengan confusion matrix, dan analisis hasil.

Data yang digunakan dalam penelitian ini adalah abstrak karya tulis ilmiah bidang psikologi yang diperoleh dari Google Scholar, terdiri dari 25 pasang abstrak yang dikelompokkan berdasarkan tipologi kemiripannya, yang meliputi copy-paste murni, parafrase kuat, mosaik, topik berbeda jauh, dan topik mirip tetapi tujuan/metode berbeda. Setiap kategori

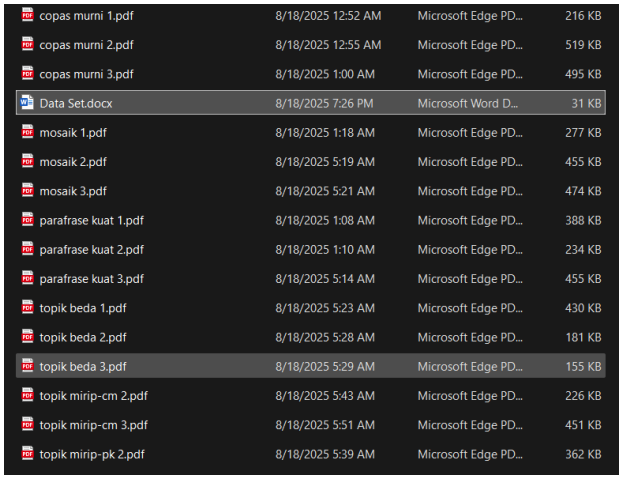
akan diuji dengan 5 pasang abstrak yang disimpan dalam format .csv, dan akan diberi label aktual menggunakan Copyleaks sebagai pembanding. Proses preprocessing bertujuan untuk memastikan keseragaman struktur teks abstrak, yang mencakup penyamaan bentuk karakter, penyederhanaan alur teks, dan perapihan spasi. Setelah preprocessing, setiap abstrak akan di-encode menjadi representasi vektor numerik dengan menggunakan Sentence-BERT. Hasil dari model Sentence-BERT akan dihitung cosine similarity untuk menentukan tingkat kemiripan antara dokumen. Evaluasi model dilakukan menggunakan confusion matrix untuk menggambarkan performa deteksi kemiripan dan classification report yang mencakup metrik seperti accuracy, precision, recall, dan F1-score. Evaluasi ini membandingkan hasil prediksi model dengan data ground truth dari Copyleaks untuk mengukur kinerja model dalam mendeteksi kemiripan antar teks.

HASIL DAN PEMBAHASAN

Hasil Pengumpulan Dataset

Pemasangan dan Pengkategorian Data

Dataset dalam penelitian ini dikumpulkan secara mandiri melalui Google Scholar. Dataset yang dikumpulkan berupa karya tulis ilmiah berbahasa Indonesia dalam bidang psikologi dan memiliki topik serta tujuan yang berbeda-beda. Total ada 25 karya tulis ilmiah yang dikumpulkan, lalu dibuatkan 5 kategori yang masing-masing berisi 5 karya tulis ilmiah.



Gambar 1. Dataset karya tulis ilmiah

Gambar 1 merupakan hasil dari pengumpulan dataset yang diambil melalui Google Scholar. Semua dataset ini diambil abstraknya saja lalu dimasukkan ke dalam 5 kategori. Masing-masing kategori memiliki 5 dataset untuk mewakilinya sebagai berikut.

1) Copy-Paste murni, dataset dalam kategori ini akan disalin dan diedit secara manual dengan mengganti huruf besar kecil dan tanda bacanya. Namun struktur kalimat dan penulisan masih dipertahankan seperti aslinya.
2) Parafrase kuat, dataset dalam kategori ini akan disalin dan diedit secara manual dengan mengganti sebagian besar kata menggunakan sinonim atau mengubah struktur kalimat, tetapi makna inti tetap sama. Tujuannya untuk menyamarkan kemiripan meskipun secara semantik masih sangat dekat.

3) Mosaik, dataset dalam kategori ini akan disalin sebagian saja dan ditambakan kalimat baru atau kutipan dari sumber lain. Jenis ini lebih sulit terdeteksi karena kemiripan tidak sepenuhnya linear, melainkan bercampur antara kalimat asli dan modifikasi.
4) Topik berbeda jauh, terdapat dua dataset yang topiknya sama sekali berbeda untuk dibandingkan.
5) Topik mirip tetapi tujuan/metode berbeda, terdapat dua dataset yang membahas tema umum yang sama (misalnya sama-sama tentang psikologi pendidikan), tetapi tujuan penelitian, metode yang digunakan, maupun hasil yang diperoleh berbeda. Dengan demikian, meskipun ada kedekatan topik, tidak ada indikasi plagiarisme.

Tabel 1 Kategori copy-paste murni

Abstrak a	Artikel ini membahas peran bahasa Indonesia dalam meningkatkan ketahanan mental individu dalam menghadapi tantangan hidup. Latar belakang masalahnya mencakup pentingnya bahasa Indonesia dalam membantu individu melihat sisi positif dari masalah, meredakan kecemasan, dan membangkitkan semangat untuk menghadapi tantangan. Fokus masalahnya adalah mengeksplorasi bagaimana bahasa Indonesia mempengaruhi persepsi diri, pola pikir, dan respon terhadap stres dan tekanan, serta bagaimana bahasa tersebut dapat digunakan sebagai alat untuk mengkomunikasikan pikiran, perasaan, dan strategi untuk meningkatkan resiliensi. Artikel ini menggunakan metode penelitian Studi literatur kualitatif. Temuan dari artikel ini mencakup analisis hubungan bahasa Indonesia dengan resiliensi yang mempengaruhi ketahanan mental remaja dalam menghadapi tantangan hidup. Beberapa poin temuan meliputi fungsi bahasa sebagai alat komunikasi, mencerminkan citra pikiran, dan kepribadian, serta peran bahasa menjembatani resiliensi dalam menghadapi tantangan hidup. Simpulan dari artikel ini menegaskan bahwa bahasa Indonesia memainkan peran penting dalam membangun ketahanan mental individu, serta pentingnya meningkatkan pemahaman tentang peran bahasa dalam membangun ketahanan mental individu untuk meningkatkan kesejahteraan psikologis masyarakat.
Abstrak b	Perilaku bullying merupakan tindakan kekerasan yang sengaja maupun tidak sengaja dilakukan oleh seseorang ataupun sekelompok baik secara verbal maupun fisik. Penelitian ini bertujuan untuk mengetahui perilaku bullying yang terjadi dan dampak psikologis apa yang dialami oleh korban bullying di sekolah dasar muhammadiyah. Metode penelitian yang digunakan adalah studi kasus kualitatif. Data diperoleh dari sumber data primer yaitu guru, kepala sekolah dan siswa selaku korban sedangkan sumber data sekunder diperoleh dari jurnal dan buku. Data dikumpulkan melalui wawancara dan observasi. Setelah data dikumpulkan data akan dianalisis menggunakan miles dan huberman yaitu pengumpulan data, reduksi data, penyajian data, dan kesimpulan. Hasil penelitian ini adalah perilaku bullying yang terjadi yaitu bullying fisik dan bullying verbal. Dampak bullying secara psikologis terlihat bahwa siswa menjadi tidak percaya diri, khawatir dengan lingkungan sekitar, trauma untuk berteman kembali, malu dengan berbicara pelan dan menghindari kontak mata, dan marah jika sudah tidak bisa dibisa menerima perlakuan buruk terus menerus.

Tabel 2. Kategori parafrase kuat

Abstrak a	Ada dampak positif dan negatif dari penggunaan teknologi terhadap pertumbuhan dan perkembangan anak. Pengguna akan menjadi kecanduan perangkat mereka jika digunakan terus-menerus. Dampak negatifnya sangat memprihatinkan karena begitu meluas di kalangan anak-anak yang menggunakan perangkat elektronik. Oleh karena itu, penting bagi orang tua untuk menyediakan, memantau, dan membatasi akses anak-anaknya terhadap perangkat elektronik. Tujuan dari penelitian ini adalah untuk mengetahui bagaimana teknologi mempengaruhi pertumbuhan mental siswa sekolah dasar. Penelitian ini bersifat kualitatif, menggunakan prosedur seperti wawancara terorganisir. Temuan penelitian ini menunjukkan bahwa perkembangan emosional anak-anak dapat dipengaruhi secara negatif oleh penggunaan perangkat elektronik yang berlebihan.

Abstrak b	Penggunaan teknologi memberikan konsekuensi ganda, baik positif maupun negatif, terhadap perkembangan anak. Jika tidak dibatasi, anak dapat mengalami kecanduan perangkat digital yang berdampak buruk pada keseharian mereka. Efek negatif ini semakin memprihatinkan karena marak terjadi pada anak usia sekolah dasar. Oleh sebab itu, peran orang tua sangat penting dalam menyediakan fasilitas, mengawasi, serta membatasi penggunaan perangkat oleh anak. Penelitian ini bertujuan untuk mengkaji pengaruh teknologi terhadap aspek mental anak sekolah dasar dengan metode kualitatif melalui wawancara terstruktur. Hasil penelitian mengungkapkan bahwa penggunaan perangkat elektronik secara berlebihan dapat mengganggu kestabilan emosi dan perkembangan psikologis anak.

Tabel 3. Kategori mosaik

Abstrak a	Media sosial telah menjadi bagian dari kehidupan sehari-hari. Seperti Penelitian telah banyak dilakukan untuk mengkaji dampak dari media sosial bagi penggunanya termasuk dampak psikologis.. Penelitian ini merupakan studi literatur dengan menggunakan metode penelitian systematic review yang bertujuan untuk menganalisis dampak psikologis pada pengguna media sosial. Data penelitian merupakan artikel penelitian yang membahas terkait dampak psikologis pada pengguna media sosial di Indonesia yang diterbitkan di Jurnal Nasional pada kurun waktu 2013 hingga tahun 2020. Pengumpulan data penelitian dilakukan melalui pencarian data di google scholar, garuda (Garba Rujukan Digital), DOAJ, dengan menyertakan kata kunci dari Media sosial yaitu 'sosial media', disertai kata kunci dari dampak psikologis ialah 'dampak psikologi'. Peneliti melakukan pengecekan antar pustaka dan membaca ulang pustaka dalam upaya menjaga ketepatan pengkajian dan mencegah kesalahan informasi dalam analisis data. Hasil dari penelitian ini diperoleh dampak psikologis pada pengguna media sosial dapat memberikan dampak positif dan negatif terhadap kesejahteraan psikologis. Dampak positif berupa dukungan sosial, mengurangi kesepian, dan rasa malu, modal social, kepekaan sosial dan emosional. Dampak negatif terjadi jika terdapat informasi berlebihan yang dapat menyebabkan 'penularan emosi' sehingga pengguna media sosial mengalami peningkatan efek psikologis negatif, timbulnya masalah signifikan dengan orang, penundaan, manajemen waktu yang buruk, dan kurang mampu mengontrol diri terhadap penggunaan jejaring sosial. Terdapat beberapa risiko yang signifikan, tetapi tidak jelas apakah risiko tersebut memengaruhi beberapa kelompok orang lebih dari yang lain, bagaimana mereka berinteraksi dengan kerentanan yang ada dan bagaimana mereka berkembang dari waktu ke waktu. Demikian pula, jelas ada peluang untuk menggunakan media sosial untuk mendukung kesejahteraan, mengurangi risiko, dan menawarkan bantuan kepada mereka yang membutuhkannya.
Abstrak b	Media sosial kini menjadi bagian penting dalam kehidupan sehari-hari dan banyak penelitian telah dilakukan untuk memahami dampak psikologisnya. Studi ini menggunakan metode *systematic review* untuk menganalisis literatur yang diterbitkan di Indonesia antara tahun 2013 hingga 2020. Data dikumpulkan melalui pencarian artikel di Google Scholar, Garuda, dan DOAJ dengan kata kunci 'sosial media' serta 'dampak psikologi'. Peneliti melakukan seleksi ketat dan membaca ulang pustaka guna memastikan keakuratan analisis. Hasil kajian menunjukkan bahwa media sosial memiliki dampak ganda terhadap kesejahteraan psikologis. Dampak positif terlihat pada dukungan sosial, pengurangan rasa kesepian, dan peningkatan sensitivitas emosional. Namun, dampak negatif juga muncul, seperti informasi berlebihan yang memicu penularan emosi, masalah relasi sosial, gangguan manajemen waktu, serta kesulitan mengendalikan penggunaan platform digital. Penelitian ini menegaskan bahwa meskipun ada risiko signifikan, media sosial juga dapat dimanfaatkan untuk meningkatkan kesejahteraan psikologis bila digunakan secara tepat.

Tabel 4. Kategori topik berbeda

Abstrak a	Dalam penelitian ini kami menyimpulkan bahwa diskriminasi atau kekerasan seksual adalah mereka yang menyakiti para pihak dengan kata-kata atau perbuatan pemaksaan, intimidasi, penahanan, tekanan psikologis, atau dengan penyalahgunaan kekuasaan atau

	penyalahgunaan lingkungan seseorang yang obsesif dan tidak mampu memberikan persetujuan yang sebenarnya itu adalah tindakan kriminal yang harus diadili. Studi ini menganalisis korban kekerasan seksual. Serta studi ini diklasifikasikan sebagai anak yang masih dilindungi oleh hukum, atau masih dibawah umur. Latar belakang studi ini yaitu meningkatnya insiden kekerasan seksual atau diskriminasi yang terjadi dalam masyarakat. Jumlah korban semakin meningkat dari tahun ketahun. Tidak ada pencegah bagi penjahat dan masyarakat untuk menghalangi mereka akan terus melakukannya.
Abstrak b	Ada dampak positif dan negatif dari penggunaan teknologi terhadap pertumbuhan dan perkembangan anak. Pengguna akan menjadi kecanduan perangkat mereka jika digunakan terus-menerus. Dampak negatifnya sangat memprihatinkan karena begitu meluas di kalangan anak-anak yang menggunakan perangkat elektronik. Oleh karena itu, penting bagi orang tua untuk menyediakan, memantau, dan membatasi akses anak-anaknya terhadap perangkat elektronik. Tujuan dari penelitian ini adalah untuk mengetahui bagaimana teknologi mempengaruhi pertumbuhan mental siswa sekolah dasar. Penelitian ini bersifat kualitatif, menggunakan prosedur seperti wawancara terorganisir. Temuan penelitian ini menunjukkan bahwa perkembangan emosional anak-anak dapat dipengaruhi secara negatif oleh penggunaan perangkat elektronik yang berlebihan.

Tabel 5. Kategori topik mirip tapi tujuan/metode berbeda

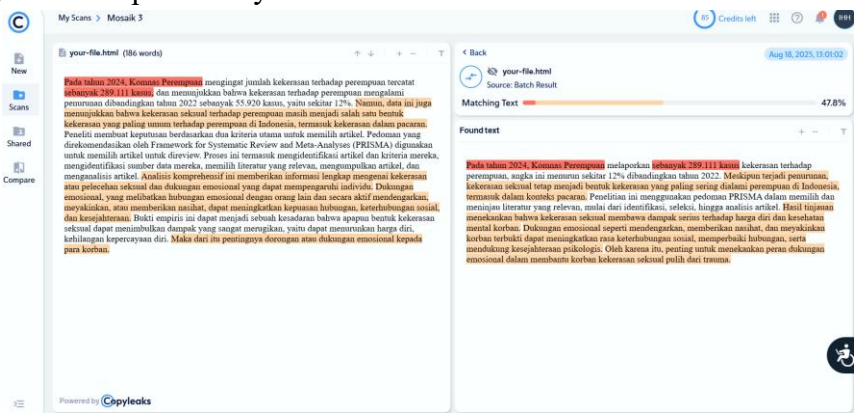
Abstrak a	Penelitian ini bertujuan untuk mengetahui hubungan Intensitas Penggunaan Gadget dengan Interaksi Sosial. Hipotesis yang diajukan adalah ada hubungan antara Intensitas penggunaan Gadget dengan Interaksi Sosial, Subjek dalam penelitian ini yakni remaja usia 13 sampai dengan 18 tahun yang tinggal di Desa Gondangmanis Kecamatan Bandarkerdungmulyo yang berjumlah 60 orang, Teknik pengambilan data yakni menggunakan Proposiv Random Sampling. Metode pengumpulan data menggunakan dua buah skala yakni skala Intensitas Penggunaan Gadget dengan Interaksi Sosial yang telah disusun oleh peneliti dan diuji validitas empiris dan reliabilitas. Analisis data dilakukan dengan metode korelasi Rho Spearman menggunakan teknik Statistical Package for Social Science (SPSS) versi 16.0. Hasil analisis data menghasilkan koefisien korelasi S'rho= -0,330 dengan p =0,010 (p< 0,05) yang berati terdapat hubungan negatif yang signifikan antara Intensitas Penggunaan Gadget dengan Interaksi Sosial yang artinya semakin tinggi Intensitas penggunaan Gadget maka semakin rendah Interaksi Sosial. Dan sebaliknya jika semakin rendah Intensitas Penggunaan Gadget maka semakin tinggi Interaksi Sosial. Jadi hipotesis yang diajukan oleh peneliti ini diterima.
Abstrak b	Ada dampak positif dan negatif dari penggunaan teknologi terhadap pertumbuhan dan perkembangan anak. Pengguna akan menjadi kecanduan perangkat mereka jika digunakan terus-menerus. Dampak negatifnya sangat memprihatinkan karena begitu meluas di kalangan anak-anak yang menggunakan perangkat elektronik. Oleh karena itu, penting bagi orang tua untuk menyediakan, memantau, dan membatasi akses anak-anaknya terhadap perangkat elektronik. Tujuan dari penelitian ini adalah untuk mengetahui bagaimana teknologi mempengaruhi pertumbuhan mental siswa sekolah dasar. Penelitian ini bersifat kualitatif, menggunakan prosedur seperti wawancara terorganisir. Temuan penelitian ini menunjukkan bahwa perkembangan emosional anak-anak dapat dipengaruhi secara negatif oleh penggunaan perangkat elektronik yang berlebihan.

Pada Tabel 1 sampai Tabel 3 memperlihatkan dataset yang sudah diedit manual, abstrak A merupakan abstrak asli dari karya tulis ilmiah. Abstrak B dalam kategori copy-paste murni, parafrase dan mosaik merupakan hasil edit dari abstrak A. Sedangkan pada Tabel 4 sampai 5 abstrak B dalam kategori topik berbeda jauh dan topik mirip tetapi tujuan/metodenya berbeda

merupakan abstrak dari dokumen A dengan kategori yang berbeda. Totalnya ada 25 pasang abstrak yang diuji.

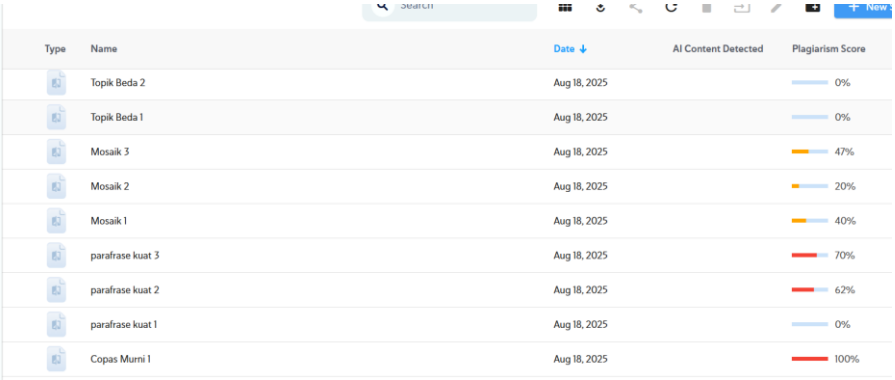
Menentukan Ground Truth

Pada tahap ini, abstrak yang sudah dipasangkan akan dibandingkan di Copyleaks untuk diketahui tingkat kemiripan teksnya.



Gambar 2. Contoh abstrak yang dibandingkan di Copyleaks

Gambar 2 merupakan salah satu contoh abstrak dari kategori mosaik yang dibandingkan. Copyleaks menyatakan tingkat kemiripan teks antara dokumen A dengan dokumen B sebesar 47,8%, nilai inilah yang nantinya akan dijadikan ground truth untuk masing-masing pasangan abstrak.



Gambar 3 Hasil pengecekan kemiripan teks dari Copyleaks

Gambar 3 menunjukkan hasil dari pengecekan pasangan abstrak. Hasil persentase dari Copyleaks akan dipetakan menjadi beberapa kelas yang terdiri dari: (1) tidak plagiarisme (≤0%), (2) plagiarisme ringan (0-<30%), (3) plagiarisme sedang (30-70%), (4) plagiarisme berat (>70%).

Tabel 6. Hasil tes dari Copyleaks

Kategori	Hasil Tes	Label
***Copy-Paste* murni 1**	100%	Plagiarisme Berat
***Copy-Paste* murni 2**	100%	Plagiarisme Berat
***Copy-Paste* murni 3**	100%	Plagiarisme Berat
***Copy-Paste* murni 4**	99%	Plagiarisme Berat
***Copy-Paste* murni 5**	100%	Plagiarisme Berat
Parafrase 1	0%	Tidak Plagiat
Parafrase 2	62%	Plagiarisme Sedang
Parafrase 3	70%	Plagiarisme Sedang
Parafrase 4	0%	Tidak Plagiat

Kategori	Hasil Tes	Label
Parafrase 5	0%	Tidak Plagiat
Mosaik 1	40%	Plagiarisme Sedang
Mosaik 2	20%	Plagiarisme Ringan
Mosaik 3	47%	Plagiarisme Sedang
Mosaik 4	35%	Plagiarisme Sedang
Mosaik 5	22%	Plagiarisme Ringan
Topik berbeda 1	0%	Tidak Plagiat
Topik berbeda 2	0%	Tidak Plagiat
Topik berbeda 3	0%	Tidak Plagiat
Topik berbeda 4	0%	Tidak Plagiat
Topik berbeda 5	0%	Tidak Plagiat
Topik mirip tapi beda tujuan/metode 1	0%	Tidak Plagiat
Topik mirip tapi beda tujuan/metode 2	0%	Tidak Plagiat
Topik mirip tapi beda tujuan/metode 3	0%	Tidak Plagiat
Topik mirip tapi beda tujuan/metode 4	0%	Tidak Plagiat
Topik mirip tapi beda tujuan/metode 5	0%	Tidak Plagiat

Tabel 6 menunjukkan hasil dari tes Copyleaks, kolom kategori berisi pasangan abstrak yang dibandingkan. Kolom hasil tes berisi persentase dari masing-masing dokumen. Pada kolom label berisi status dari pasangan abstrak berdasarkan persentase di kolom hasil tes. Jika hasilnya ≤0% maka diberi label tidak plagiarisme, jika skornya 0-<30% maka diberi label plagiarisme ringan, jika skornya 30-70% maka diberi label plagiarisme sedang, dan jika skornya >70% maka diberi label plagiarisme berat.

Dari hasil tes ini terdapat 13 pasang abstrak yang terdeteksi tidak plagiat, 2 pasang abstrak yang terdeteksi plagisme ringan, 5 pasang abstrak yang terdeteksi plagiarism sedang, dan 5 pasang abstrak yang terdeteksi plagiarisme berat.

Tahap Preprocessing Data

Pada tahap ini, data berupa abstrak penelitian dipersiapkan agar dapat diproses lebih lanjut oleh model. Proses preprocessing sederhana dilakukan untuk memastikan data berada pada format yang seragam dan bersih dari komponen yang tidak diperlukan. Tahapan preprocessing meliputi menyamakan bentuk karakter, membuat alur teks menjadi lurus satu baris, dan merapikan spasi sebelum tanda baca.

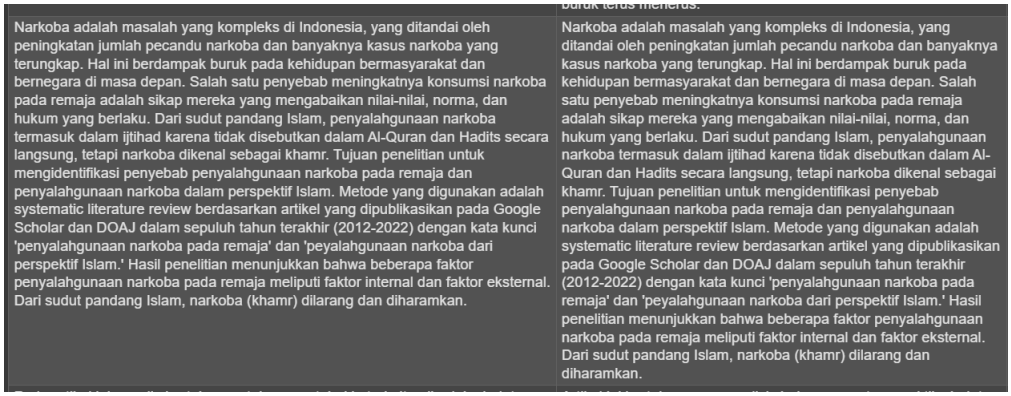
```python
def normalize_unicode(s: str) -> str:
    # samakan bentuk karakter (kutip miring/dash → bentuk umum)
    return unicodedata.normalize('NFKC', s)

def normalize_punct(s: str) -> str:
    # ganti em/en dash ke hyphen, rapikan spasi sebelum tanda baca
    s = s.replace('—', '-').replace('–', '-')
    s = re.sub(r'\s+([,.;:!?])', r'\1', s)   # hapus spasi sebelum tanda baca
    s = re.sub(r'([,.;:!?])([^\s])', r'\1 \2', s)  # pastikan ada spasi setelah tanda baca
    s = re.sub(r'\s+', ' ', s)  # collapse whitespace
    return s.strip()

def remove_ctrl(s: str) -> str: #bikin abstrak jadi lurus, tidak ada baris baru
    return ''.join(ch for ch in s if ch.isprintable())
```

Gambar 4. Kode fungsi preprocessing

Gambar 4 merupakan kode program untuk melakukan preprocessing data. Ketiganya akan dipanggil melalui clean_texs(). Hasilnya dapat dilihat pada Gambar 4.5.



Narkoba adalah masalah yang kompleks di Indonesia, yang ditandai oleh peningkatan jumlah pecandu narkoba dan banyaknya kasus narkoba yang terungkap. Hal ini berdampak buruk pada kehidupan bermasyarakat dan bernegara di masa depan. Salah satu penyebab meningkatnya konsumsi narkoba pada remaja adalah sikap mereka yang mengabaikan nilai-nilai, norma, dan hukum yang berlaku. Dari sudut pandang Islam, penyalahgunaan narkoba termasuk dalam ijtihad karena tidak disebutkan dalam Al-Quran dan Hadits secara langsung, tetapi narkoba dikenal sebagai khamr. Tujuan penelitian untuk mengidentifikasi penyebab penyalahgunaan narkoba pada remaja dan penyalahgunaan narkoba dalam perspektif Islam. Metode yang digunakan adalah systematic literature review berdasarkan artikel yang dipublikasikan pada Google Scholar dan DOAJ dalam sepuluh tahun terakhir (2012-2022) dengan kata kunci 'penyalahgunaan narkoba pada remaja' dan 'peyalahgunaan narkoba dari perspektif Islam.' Hasil penelitian menunjukkan bahwa beberapa faktor penyalahgunaan narkoba pada remaja meliputi faktor internal dan faktor eksternal. Dari sudut pandang Islam, narkoba (khamr) dilarang dan diharamkan.	Narkoba adalah masalah yang kompleks di Indonesia, yang ditandai oleh peningkatan jumlah pecandu narkoba dan banyaknya kasus narkoba yang terungkap. Hal ini berdampak buruk pada kehidupan bermasyarakat dan bernegara di masa depan. Salah satu penyebab meningkatnya konsumsi narkoba pada remaja adalah sikap mereka yang mengabaikan nilai-nilai, norma, dan hukum yang berlaku. Dari sudut pandang Islam, penyalahgunaan narkoba termasuk dalam ijtihad karena tidak disebutkan dalam Al-Quran dan Hadits secara langsung, tetapi narkoba dikenal sebagai khamr. Tujuan penelitian untuk mengidentifikasi penyebab penyalahgunaan narkoba pada remaja dan penyalahgunaan narkoba dalam perspektif Islam. Metode yang digunakan adalah systematic literature review berdasarkan artikel yang dipublikasikan pada Google Scholar dan DOAJ dalam sepuluh tahun terakhir (2012-2022) dengan kata kunci 'penyalahgunaan narkoba pada remaja' dan 'peyalahgunaan narkoba dari perspektif Islam.' Hasil penelitian menunjukkan bahwa beberapa faktor penyalahgunaan narkoba pada remaja meliputi faktor internal dan faktor eksternal. Dari sudut pandang Islam, narkoba (khamr) dilarang dan diharamkan.

Gambar 5 Hasil dari preprocessing

Gambar 5 menunjukkan abstrak yang telah melewati tahap prerpocessing, pada tahap ini data sudah tidak memiliki spasi yang berganda dan tidak memiliki spasi di awal dan akhir kalimat.

Implementasi Sentence-BERT

Pada tahap ini, mengubah teks abstrak menjadi representasi numerik menggunakan Sentence-BERT. Sentence-BERT merupakan pengembangan dari BERT yang dioptimalkan untuk menghasilkan sentence embedding, sehingga lebih efektif dalam menghitung kesamaan antar kalimat atau dokumen pendek.

Implementasi dilakukan dengan memanggil pustaka sentence-transformers menggunakan model pre-trained distiluse-base-multilingual-cased-v1. Setiap abstrak kemudian dikonversi menjadi vektor berdimensi 512. Proses ini memungkinkan komputer untuk memahami makna semantik dari teks abstrak.

1. Pemanggilan Model

Langkah pertama adalah memanggil model pre-trained Sentence-BERT melalui library sentence-transformers. Dengan perintah pada Gambar 4.6.

```
# ===== Load model SBERT =====
model = SentenceTransformer(MODEL_NAME)
```

Gambar 6. Kode proses

Model akan terinstansiasi dan siap digunakan untuk mengubah teks menjadi embedding berdimensi tinggi.

2. Proses Encoding Abstrak

Dua kolom teks utama dalam dataset, yaitu abstrak_a dan abstrak_b, diubah menjadi representasi numerik (embedding vector) menggunakan fungsi encode(). Agar lebih efisien saat menangani banyak data, proses ini dilakukan secara batch dengan ukuran 32 data per proses. Parameter tambahan seperti convert_to_tensor=False digunakan untuk menghasilkan output dalam bentuk array NumPy, sedangkan show_progress_bar=True digunakan agar proses dapat dipantau. Kode yang digunakan terdapat pada Gambar 7.

```
# ===== Encode & similarity =====
emb_a = model.encode(
    df['abstrak_a'].tolist(),
    batch_size=32,
    convert_to_tensor=False,
    show_progress_bar=True
    )
emb_b = model.encode(
    df['abstrak_b'].tolist(),
    batch_size=32,
    convert_to_tensor=False,
    show_progress_bar=True
    )
```

Gambar 7. Kode encode

Output dari proses ini adalah dua buah matriks vektor (emb_a dan emb_b) yang masing-masing merepresentasikan teks dari abstrak A dan abstrak B.

```
Contoh embedding (pair_id=1) – 12 elemen pertama:
emb_a[0][:12] = [ 0.032677  0.028961  0.00705   0.000365 -0.020723  0.022197 -0.056154
  0.050476 -0.013697  0.037291  0.001461 -0.04306 ]
emb_b[0][:12] = [ 0.032677  0.028961  0.00705   0.000365 -0.020723  0.022197 -0.056154
  0.050476 -0.013697  0.037291  0.001461 -0.04306 ]
```

Gambar 8. Hasil dari model Sentence-BERT

Gambar 8 menunjukkan sebagian hasil dari Sentence-BERT berupa embedding dari abstrak A dan abstrak B. Pada abstrak A, 12 elemen pertama menghasilkan emb_a[0][:12] = [0.032677 0.028961 0.00705 0.000365 -0.020723 0.022197 -0.056154 0.050476 -0.013697 0.037291 0.001461 -0.04306], dan pada abstrak B, 12 elemen pertama menghasilkan emb_b[0][:12] = [0.032677 0.028961 0.00705 0.000365 -0.020723 0.022197 -0.056154 0.050476 -0.013697 0.037291 0.001461 -0.04306].

Menghitung Kesamaan dengan Cosine Similarity

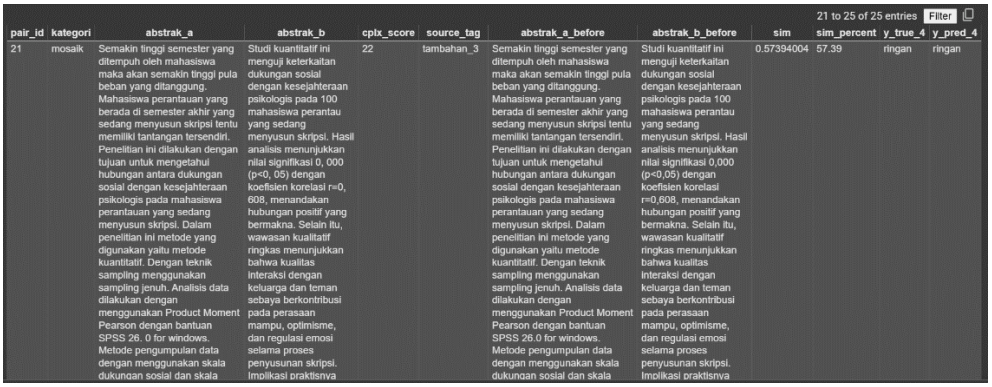
Setelah teks direpresentasikan dalam bentuk vektor, tahap berikutnya adalah menghitung tingkat kesamaan antar pasangan abstrak. Perhitungan ini dilakukan menggunakan metode cosine similarity, yang mampu mengukur kedekatan semantik antar dua vektor.

```
num = np.sum(emb_a * emb_b, axis=1)
den = (np.linalg.norm(emb_a, axis=1) * np.linalg.norm(emb_b, axis=1))
sim = num / np.maximum(den, 1e-12)
df['sim'] = sim
df['sim_percent'] = (df['sim'] * 100).round(2)
```

Gambar 9. Kode untuk metode cosine similarity

Pada Gambar 9 merupakan kode untuk menjalankan metode cosine similarity. Tiap pasangan (Ai, Bi) dibandingkan tingkat kemiripannya menggunakan cosine similarity. Program menghitung pembilang berupa hasil kali dalam (dot product) antara vektor Ai dan Bi untuk setiap baris (disimpan pada variabel num). Selanjutnya, program menghitung penyebut sebagai hasil kali panjang vector, masing-masing embedding (disimpan pada den). Nilai cosine untuk setiap pasangan kemudian diperoleh dari pembilang dibagi penyebut. Untuk menjaga stabilitas numerik dan mencegah pembagian dengan nol, penyebut diberi batas bawah 10-12. Hasil akhirnya disimpan sebagai sim (rentang [−1,1]) dan juga disajikan dalam bentuk persentase (sim_percent) agar mudah dibaca pada tabel hasil.

Menentukan apakah sebuah pasangan abstrak dianggap sebagai plagiat atau tidak, digunakan nilai ambang batas (threshold) sebesar 50%. Apabila nilai kesamaan (similarity score) lebih besar atau sama dengan 0.50 (50%), maka pasangan teks dikategorikan sebagai plagiat (pred = 1). Sebaliknya, jika nilai kesamaan di bawah ambang batas tersebut, maka pasangan teks dianggap tidak plagiat (pred = 0).



21 to 25 of 25 entries Filter

pair_id	kategori	abstrak_a	abstrak_b	cplx_score	source_tag	abstrak_a_before	abstrak_b_before	sim	sim_percent	y_true_4	y_pred_4
21	mosaik	Semakin tinggi semester yang ditempuh oleh mahasiswa maka akan semakin tinggi pula beban yang ditanggung. Mahasiswa perantauan yang berada di semester akhir yang sedang menyusun skripsi tentu memiliki tantangan tersendiri. Penelitian ini dilakukan dengan tujuan untuk mengetahui hubungan antara dukungan sosial dengan kesejahteraan psikologis pada mahasiswa perantauan yang sedang menyusun skripsi. Dalam penelitian ini metode yang digunakan yaitu metode kuantitatif. Dengan teknik sampling menggunakan sampling jenuh. Analisis data dilakukan dengan menggunakan Product Moment Pearson dengan bantuan SPSS 26. 0 for windows. Metode pengumpulan data dengan menggunakan skala dukungan sosial dan skala	Studi kuantitatif ini menguji keterkaitan dukungan sosial dengan kesejahteraan psikologis pada 100 mahasiswa perantau yang sedang menyusun skripsi. Hasil analisis menunjukkan nilai signifikasi 0, 000 (p<0, 05) dengan koefisien korelasi r=0, 608, menandakan hubungan positif yang bermakna. Selain itu, wawasan kualitatif ringkas menunjukkan bahwa kualitas interaksi dengan keluarga dan teman sebaya berkontribusi pada perasaan mampu, optimisme, dan regulasi emosi selama proses penyusunan skripsi. Implikasi praktisnya	22	tambahan_3	Semakin tinggi semester yang ditempuh oleh mahasiswa maka akan semakin tinggi pula beban yang ditanggung. Mahasiswa perantauan yang berada di semester akhir yang sedang menyusun skripsi tentu memiliki tantangan tersendiri. Penelitian ini dilakukan dengan tujuan untuk mengetahui hubungan antara dukungan sosial dengan kesejahteraan psikologis pada mahasiswa perantauan yang sedang menyusun skripsi. Dalam penelitian ini metode yang digunakan yaitu metode kuantitatif. Dengan teknik sampling menggunakan sampling jenuh. Analisis data dilakukan dengan menggunakan Product Moment Pearson dengan bantuan SPSS 26.0 for windows. Metode pengumpulan data dengan menggunakan skala dukungan sosial dan skala	Studi kuantitatif ini menguji keterkaitan dukungan sosial dengan kesejahteraan psikologis pada 100 mahasiswa perantau yang sedang menyusun skripsi. Hasil analisis menunjukkan nilai signifikasi 0,000 (p<0,05) dengan koefisien korelasi r=0,608, menandakan hubungan positif yang bermakna. Selain itu, wawasan kualitatif ringkas menunjukkan bahwa kualitas interaksi dengan keluarga dan teman sebaya berkontribusi pada perasaan mampu, optimisme, dan regulasi emosi selama proses penyusunan skripsi. Implikasi praktisnya	0.57394004	57.39	ringan	ringan

Gambar 10. Hasil dari perhitungan cosine similarity

Pada kolom abstak_a dan abstrak_b berisi abstrak yang dibandingkan. Kolom cplx_score berisi penilaian dari Copyleaks , nilainya berupa persentase (0–100) yang akan dipakai sebagai label acuan apakah dua teks dianggap plagiat atau tidak. Kolom kategori berisi kategori dari setiap pasang dokumen, pada penelitian ini terdapat 5 kategori, yaitu copy-paste murni, parafrase kuat, mosaik, topik berbeda jauh, dan topik mirip tapi tujuan/metode berbeda. Pada kolom sim memperlihatkan hasil deteksi kemiripan dari model Sentence-BERT menggunakan cosine similarity dengan nilai antara 0 sampai 1. Semakin mendekati angka 1, maka kedua abstrak dikatakan semakin mirip dan semakin mendekati angka 0, maka kedua abstrak dikatakan semakin berbeda. Sim_percent adalah konversi persentase dari kolom sim yang dibulatkan ke dua angka desimal. Pada kolom y_true_4 berisi status prediksi dari Copyleaks berdasarkan tingkatan plagiarisme jika hasilnya ≤0% maka diberi label tidak, jika skornya 0-<30% maka diberi label ringan, jika skornya 30-70% maka diberi label sedang, dan jika skornya >70% maka diberi label berat.. Pada kolom y_pred_4 berisi prediksi dari model Sentence-BERT berdasarkan tingkatan plagiarisme sama seperti pada kolom y_true_4.

A. Evaluasi Model

1) Confusion Matrix

Pada tahap ini, hasil dari perthitungan cosine similarity akan dibandingkan dengan ground truth dari Copyleaks dengan memasukannya confusion matrix. Pada penelitian kode ditulis sebagai berikut.

```
labels4 = ["tidak","ringan","sedang","berat"]
cm4 = confusion_matrix(df['y_true_4'], df['y_pred_4'], labels=labels4)
cm4_df = pd.DataFrame(cm4, index=[f"GT_{l}" for l in labels4], #baris
                           columns=[f"P_{l}"  for l in labels4]) #kolom
print("\nConfusion Matrix 4x4:\n", cm4_df)
```

Gambar 11. Kode untuk Confusion Matrix

Gambar 11 merupakan kode untuk menggunakan Confusion Matrix. y_true_4 berisi ground truth dari Copyleaks dan y_pred_4 berisi hasil prediksi dari model Sentence-BERT.

Tabel 7. Hasil confusion matrix

	P_tidak	P_ringan	P_sedang	P_berat
GT_tidak	1	8	3	1
GT_ringan	0	1	0	1
GT_sedang	0	0	0	5
GT_berat	0	0	0	5

Pada Tabel 7 memperlihatkan hasil dari confusion matrix yang dilakukan dalam bentuk 4 × 4. Kolom P_tidak, P_ringan, P_sedang, dan P_berat berisi data prediksi dari model Sentence_BERT. Sedangkan kolom GT_tidak, GT_ringan, GT_sedang dan GT_berat berisi data aktual (ground truth) dari Copyleaks, bisa disimpulkan sebagai berikut:

a. Pada kelas tidak:
   TN = 12 (data aktual tidak plagiat → diprediksi tidak plagiat)
   FP = 0 (data aktual tidak plagiat → tapi diprediksi plagiat)
   FN = 12 (data aktual plagiat → tapi diprediksi tidak plagiat)
   TP = 1 (data aktual plagiat → diprediksi plagiat)
b. Pada kelas ringan:
   TN = 15 (data aktual tidak plagiat → diprediksi tidak plagiat)
   FP = 8 (data aktual tidak plagiat → tapi diprediksi plagiat)
   FN = 1 (data aktual plagiat → tapi diprediksi tidak plagiat)
   TP = 1 (data aktual plagiat → diprediksi plagiat)
c. Pada kelas sedang:
   TN = 17 (data aktual tidak plagiat → diprediksi tidak plagiat)
   FP = 3 (data aktual tidak plagiat → tapi diprediksi plagiat)
   FN = 5 (data aktual plagiat → tapi diprediksi tidak plagiat)
   TP = 0 (data aktual plagiat → diprediksi plagiat)
d. Pada kelas berat:
   TN = 13 (data aktual tidak plagiat → diprediksi tidak plagiat)
   FP = 7 (data aktual tidak plagiat → tapi diprediksi plagiat)
   FN = 0 (data aktual plagiat → tapi diprediksi tidak plagiat)
   TP = 5 (data aktual plagiat → diprediksi plagiat)

Hasil evaluasi menunjukkan kecenderungan model untuk menaikkan prediksi ke tingkat keparahan yang lebih tinggi. Pola ini tercermin dari keberhasilan model menangkap seluruh kasus plagiarisme berat (TP=5, FN=0), namun diikuti oleh jumlah alarm palsu yang relatif tinggi pada kelas tersebut (FP=7) yang sebagian besar berasal dari kelas sedang. Dengan demikian, model berorientasi pada sensitivitas terhadap kasus berat dengan konsekuensi spesifisitas yang menurun pada kelas non-berat.

1. Kelas berat: seluruh kasus berat terdeteksi (TP=5; FN=0), menandakan risiko kelolosan kasus berat dapat ditekan. Namun, adanya FP=7 menunjukkan overestimasi ke kelas berat, terutama terhadap kasus yang sebenarnya berada pada kelas sedang.
2. Kelas sedang: tidak ada kasus sedang yang dikenali tepat (TP=0; FN=5), seluruhnya berpindah ke prediksi “berat”. Hal ini mengindikasikan batas pemisah antara “sedang” dan “berat” yang terlalu agresif.

3. Kelas ringan: dari dua kasus ringan, satu teridentifikasi benar (TP=1; FN=1), tetapi prediksi "ringan" muncul sembilan kali (FP=8), menandakan ketidakstabilan label ringan pada ukuran data saat ini.
4. Kelas tidak: hanya 1 dari 13 kasus yang dikenali sebagai "tidak" (TP=1; FN=12), sementara model hampir tidak pernah mengeluarkan label "tidak" (FP=0 karena total prediksi "tidak" hanya satu). Ini menunjukkan model reluktan menetapkan "tidak" dan cenderung mengeskalasi ke level di atasnya.

2) Classification Report

Classification report atau hasil laporan klasifikasi pada sistem yang terdiri dari nilai-nilai precision, recall, accuracy, dan f1-score dirincikan pada Tabel 8.

Tabel 8. Hasil classification report

Keterangan	*precision*	*recall*	*f1-score*	*support*
tidak	1.000	0.077	0.143	13
ringan	0.111	0.500	0.182	2
sedang	0.000	0.000	0.000	5
berat	0.417	1.000	0.588	5
Accuracy			0.280	25
Macro Average	0.382	0.394	0.228	25
Weighted Average	0.612	0.280	0.206	25

Tabel 4.3 menunjukkan hasil evaluasi berdasarkan parameter evaluasi precision, recall, accuracy, dan f1-score. Pada evaluasi empat kelas, nilai kesamaan Sentence-BERT yang sudah dikonversi ke persentase (sim_percent) dipetakan ke label yang sama dengan ground truth Copyleaks: 0% = tidak, 0% < s < 30% = ringan, 30% ≤ s ≤ 70% = sedang, s > 70% = berat. Catatan: ambang 50% hanya digunakan pada skenario biner (plagiat/tidak). Berdasarkan Tabel 8, informasi ini menjadi dasar untuk menghitung tingkat akurasi sistem secara keseluruhan.

Perhitungan nilai accuracy dilakukan seperti pada persamaan (4.1):

$$\boldsymbol{Accuracy = \frac{1+1+0+5}{13+2+5+5} \times 100\%} \quad \boldsymbol{(4.1)}$$

$$\boldsymbol{Accuracy = \frac{7}{25} \times 100\%}$$

$$\boldsymbol{Accuracy = 28\%}$$

Hasil perhitungan pada persamaan (4.1) mengonfirmasi bahwa sistem deteksi kemiripan antar teks berbasis model Sentence-BERT mencapai tingkat akurasi sebesar 28% dalam mendeteksi persamaan antar teks.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut: Penerapan metode Sentence-BERT untuk mendeteksi tingkat kemiripan antar teks. Penelitian ini berhasil menerapkan pipeline deteksi kemiripan berbasis Sentence-BERT (model distiluse-base-multilingual-cased-v1) pada pasangan abstrak karya ilmiah psikologi berbahasa Indonesia. Teks dipreproses ringan, di-encode menjadi sentence embeddings, lalu dihitung kesamaannya dengan cosine similarity. Skema pelabelan prediksi kemudian dipetakan ke empat tingkat (tidak, ringan, sedang, berat) agar sejalan dengan label ground truth dari Copyleaks. Pendekatan ini dapat mendeteksi kemiripan semantik lintas parafrase sederhana tanpa bergantung pada kecocokan kata mentah, serta relatif efisien untuk evaluasi banyak pasangan dokumen. Akurasi model dalam mengidentifikasi kemiripan teks

pada karya ilmiah psikologi berbahasa Indonesia. Dengan dataset 25 pasang abstrak dan evaluasi 4 kelas, model mencapai akurasi 28%, dengan ciri performa: recall sangat tinggi pada kelas berat (seluruh kasus berat terdeteksi), namun terjadi over-prediction ke kelas berat sehingga banyak kasus sedang terangkat menjadi berat dan kelas tidak sering salah dikenali. Temuan ini menunjukkan model efektif menangkap plagiarisme eksplisit/tinggi, tetapi membutuhkan kalibrasi agar lebih presisi pada rentang tidak-ringan-sedang.

DAFTAR PUSTAKA

Ambarwati, D., Wibowo, U. B., Arsyiadanti, H., & Susanti, S. (2021). Studi literatur: Peran inovasi pendidikan pada pembelajaran berbasis teknologi digital. *Jurnal Inovasi Teknologi Pendidikan*, *8*(2), 173–184.

Azmi, M. (2022). Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan Tf-Idf. TEKNIMEDIA: Teknologi Informasi Dan Multimedia, 2(2), 90–95. https://doi.org/10.46764/teknimedia.v2i2.51

Ekaningtyas, N. L. D. (2022). Psikologi Dalam Dunia Pendidikan. Padma Sari: Jurnal Ilmu Pendidikan, 2(01), 29–38. https://doi.org/10.53977/ps.v2i01.526

Fad'li, G. A., Marsofiyati, M., & Suherdi, S. (2023). Implementasi Arsip Digital Untuk Penyimpanan Dokumen Digital. *Jurnal Manuhara: Pusat Penelitian Ilmu Manajemen Dan Bisnis*, *1*(4), 1–10.

Hanum, A. N. L., Sahidi, S., Madeten, S. S., & Amir, A. (2021). Pelatihan Manajemen Referensi: Strategi Menghindari Aksi Plagiarisme Di Kalangan Mahasiswa Menggunakan Zotero. Dharmakarya, 10(4), 307. https://doi.org/10.24198/dharmakarya.v10i4.35127

Herlambang, H., Suwita, J., & Tiara, B. (2021). Analisa Dan Perancangan Sistem Pendeteksi Plagiarisme Skripsi Pada Stmik Insan Pembangunan Menggunakan Metode Cosine Similarity. Insan Pembangunan Sistem Informasi Dan Komputer (IPSIKOM), 9(1). https://doi.org/10.58217/ipsikom.v9i1.188

Hestiani, D., & Suriyani, A. (2023). Upaya Penangan Plagiarisme di Institusi Perguruan Tinggi. Jurnal Indopedia, 1(4), 1536–1545.

Horbach, S. P. J. M. (Serge., & Halffman, W. (Willem). (2019). The extent and causes of academic text recycling or 'self-plagiarism.' Research Policy, 48(2), 492–502. https://doi.org/10.1016/j.respol.2017.09.004

Munawar, Z., Herdiana, Y., Suharya, Y., & Indah Putri, N. (2021). Pemanfaatan Teknologi Digital Di Masa Pandemi Covid-19. Tematik, 8(2), 160–175. https://doi.org/10.38204/tematik.v8i2.689

Pawestri, S., & Suyanto, Y. (2024). Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia. Jurnal Media Informatika Budidarma, 8(3), 1440. https://doi.org/10.30865/mib.v8i3.7648

Pratiwi, M. A., & Aisya, N. (2021). Fenomena plagiarisme akademik di era digital. Publishing Letters, 1(2), 16–33. https://doi.org/10.48078/publetters.v1i2.23

Purba, N. F., Annisa, F. S., Syafitri, A., & Purba, S. H. (2024). Jurnal Kesehatan Unggul Gemilang. Pemanfaatan Teknologi Digital Dalam Pelayanan Kesehatan Publik: Sebuah Tinjauan Analisis Kebijakan, 8(1), 7–15.

Salbiah, N. A. (2021). Kasus Plagiarisme di Tingkat SD hingga SMA Capai 94 Persen. https://www.jawapos.com/pendidikan/01322792/kasus-plagiarisme-di-tingkat-sd-hingga-sma-capai-94-persen

Shadiqi, M. A. (2019). Memahami dan Mencegah Perilaku Plagiarisme dalam Menulis Karya Ilmiah. Buletin Psikologi, 27(1), 30. https://doi.org/10.22146/buletinpsikologi.43058

Sinurat, H., Yunita, E., & Sumanti, R. (2021). Plagiarisme dalam budaya penulisan karya tulis ilmiah. Jurnal Transformasi Administrasi, 11(September 2021), 139–151.

Sugiarto, S., Prayitno, & Karneli, Y. (2021). Peran Psikologi Dalam Konseling. KENDURI: Jurnal Pengabdian Dan Pemberdayaan Masyarakat, 01(01), 27–30. https://siducat.org/index.php/kenduri

Susilo, Y., Wijayanti, E., & Santoso, S. (2021). Penerapan Teknologi Digital Pada Ekonomi Kreatif Pada Bisnis Minuman Boba. Jurnal Ekonomi Manajemen Sistem Informasi, 2(4), 457–468. https://doi.org/10.31933/jemsi.v2i4.383

Swari, M. H. P., Putra, C. A., & Handika, I. P. S. (2021). Plagiarism Checker pada Sistem Manajemen Data Tugas Akhir. Jurnal Sains Dan Informatika, 7(2), 192–201. https://doi.org/10.34128/jsi.v7i2.338

Wulansari, W., Fauziyah, D., Hidayat, T., Ramasiah, S., Prehanto, A., & Nuryadin, A. (2022). Perkembangan Industri Kreatif Di Kota Tasikmalaya Pada Era Digital. Jurnal Industri Kreatif Dan Kewirausahaan, 5(2), 122–129. https://doi.org/10.36441/kewirausahaan.v5i2.1313