JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Pengembangan Sistem Deteksi Plagiarisme Dokumen Jurnal Berbasis Bidirectional Encoder Representations from Transformers Dan Cosine Similarity

Cahya Yoga Ariyanto*, Adam Sekti Aji

Sains dan Teknologi, Infromatika, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
Email: [1*]cahyayoga10@gmail.com, [2]adamaji@staff.uty.ac.id
Email penulis korespondensi: cahyayoga10@gmail.com
Submitted **07-11-2025;** Accepted **11-12-2025;** Published **31-12-2025**

Abstrak

Perkembangan teknologi digital telah membawa dampak yang cukup besar dalam berbagai bidang, termasuk pendidikan dan pengelolaan dokumen ilmiah. Kemudahan akses terhadap jurnal online memunculkan tantangan baru berupa meningkatnya potensi plagiarisme. Untuk mengatasi hal tersebut, diperlukan sistem otomatis yang mampu mendeteksi kemiripan antar dokumen dengan cepat dan akurat. Penelitian ini bertujuan untuk mengembangkan sistem deteksi plagiarisme berbasis Cosine Similarity dan Bidirectional Encoder Representations from Transformers (BERT). Tahapan penelitian meliputi preprocessing teks, pembobotan kata menggunakan Term Frequency–Inverse Document Frequency (TF-IDF), perhitungan Cosine Similarity, pelatihan model BERT, serta evaluasi performa model. Hasil penelitian menunjukkan bahwa penggabungan BERT dengan TF-IDF memberikan peningkatan performa yang signifikan dibandingkan dengan penggunaan BERT murni. Berdasarkan hasil pengujian, model BERT dengan TF-IDF mencapai akurasi tertinggi sebesar 0.9621 pada skenario pembagian data 10:90, dengan nilai precision 0.8141, recall 0.7302, dan F1-score 0.8022. Sementara itu, model BERT tanpa TF-IDF hanya memperoleh akurasi sebesar 0.8529. Penerapan Cosine Similarity dengan nilai ambang batas (threshold) 0.6 juga terbukti efektif dalam mengidentifikasi dokumen plagiat dan non-plagiat. Hasil ini membuktikan bahwa kombinasi BERT dan TF-IDF dapat meningkatkan keakuratan sistem deteksi plagiarisme dengan memahami konteks semantik dan bobot kata secara bersamaan.

Kata Kunci: Plagiarisme; Cosine Similarity; BERT; TF-IDF; NLP

Abstract

The development of digital technology has had a significant impact across various fields, including education and the management of scientific documents. The ease of access to online journals has introduced a new challenge—an increase in the potential for plagiarism. To address this issue, an automated system capable of detecting document similarity quickly and accurately is required. This study aims to develop a plagiarism detection system based on Cosine Similarity and Bidirectional Encoder Representations from Transformers (BERT). The research stages include text preprocessing, word weighting using Term Frequency–Inverse Document Frequency (TF-IDF), Cosine Similarity computation, BERT model training, and model performance evaluation. The results show that integrating BERT with TF-IDF significantly improves performance compared to using BERT alone. Based on the experiments, the BERT model with TF-IDF achieved the highest accuracy of 0.9621 in a 10:90 data split scenario, with a precision of 0.8141, recall of 0.7302, and F1-score of 0.8022. Meanwhile, the BERT model without TF-IDF only achieved an accuracy of 0.8529. The application of Cosine Similarity with a threshold value of 0.6 also proved effective in identifying plagiarized and non-plagiarized documents. These findings demonstrate that combining BERT and TF-IDF enhances the accuracy of plagiarism detection systems by simultaneously capturing semantic context and word weighting.

Keywords: Plagiarisme; Cosine Similarity; BERT; TF-IDF; NLP

1. PENDAHULUAN

Perkembangan teknologi yang begitu pesat saat ini telah mengubah pola hidup manusia menuju era digital. Teknologi telah menjadi salah satu kebutuhan pokok dalam kehidupan manusia hampir di semua bidang [1]. Pada era digital saat ini, teknologi membawa dampak positif maupun negatif. Salah satu bidang yang terdampak pengaruh dari kemajuan teknologi adalah bidang pendidikan, terutama dalam pengelolaan dan penggunaan dokumen digital.

Dokumen digital merupakan salah satu hasil dari kemajuan teknologi [2], contohnya adalah jurnal online. Jurnal online menjadi bentuk dokumen digital yang memiliki peran penting dan dibutuhkan di berbagai bidang. Secara umum, jurnal adalah publikasi yang tersedia di perpustakaan dan berisi informasi, berita, serta hasil penelitian dari beragam topik. Jurnal sendiri dapat ditemukan dalam dua bentuk, yakni cetak dan non-cetak atau digital. Jurnal online merupakan versi digital dari jurnal cetak yang umumnya tersedia di perpustakaan dan dapat diakses melalui email, situs web, atau jaringan internet. Sama seperti jurnal cetak, jurnal online juga termasuk terbitan berseri, namun keduanya berbeda dari segi media penyajiannya jurnal cetak menggunakan kertas, sedangkan jurnal online dapat dibaca langsung secara daring tanpa perlu dicetak [3].

Adanya jurnal online memiliki banyak manfaat, seperti dapat dibaca langsung secara daring di mana saja kapan saja tanpa perlu membawa cetakkan dari jurnal tersebut. Namun ,dengan adanya kemudahan tersebut berpotensi muncul tindakkan plagiat jurnal. Banyak cara yang dapat dilakukan dalam palagiat, pada umumnya Adalah menyalin dan memodifikasi artikel dari jurnal online [4].

Salah satu teknik mendeteksi plagiat dokumen jurnal adalah dengan cara manual, yaitu membaca dan membandingkan beberapa jurnal secara langsung [5]. Langkah ini bertujuan untuk menemukan adanya kesamaan atau kemiripan makna yang dapat mengindikasikan praktik plagiarisme, seperti kesamaan dalam struktur kalimat, paragraf, maupun isi secara keseluruhan. Melalui proses tersebut, dapat diketahui apakah suatu jurnal benar orisinal atau hanya merupakan hasil penyalinan dari karya ilmiah yang sudah ada sebelumnya. Namun, metode manual ini memiliki

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

keterbatasan karena membutuhkan waktu yang lama, tenaga yang besar, dan sangat bergantung pada ketelitian pembaca [6]. Selain itu, tingkat keakuratannya pun tidak selalu konsisten karena faktor subjektivitas manusia [3]. Oleh sebab itu, dibutuhkan pendekatan yang lebih sistematis dan otomatis, seperti pemanfaatan teknologi komputasi dan algoritma pendeteksi kemiripan teks, agar proses verifikasi keaslian jurnal menjadi lebih cepat, *objektif*, dan akurat.

Cara yang lebih efektif untuk melakukan deteksi plagiarisme adalah dengan menggunakan sistem otomatis berbasis teknologi kecerdasan buatan (*Artificial Intelligence*) yang mampu menganalisis dan membandingkan isi dokumen jurnal secara menyeluruh [7], [8]. Sistem ini bekerja dengan memanfaatkan algoritma cosine similarity dan *Bidirectional Encoder Representations from Transformers* untuk mengidentifikasi kemiripan struktur kalimat, gaya penulisan, serta makna semantik antar dokumen.

Melalui pendekatan ini, proses deteksi tidak hanya bergantung pada pencocokan kata atau kalimat secara literal, tetapi juga mampu memahami makna dan parafrase yang terkandung didalam dokumen. Dengan demikian, sistem dapat mendeteksi plagiarisme yang bersifat langsung (*copy paste*) maupun tidak langsung (parafrase atau modifikasi kalimat) [8]. Selain itu, sistem otomatis ini dapat diintegrasikan dengan basis data jurnal ilmiah yang luas, sehingga mampu melakukan pemeriksaan lintas sumber dengan kecepatan tinggi dan tingkat akurasi yang lebih baik dibandingkan metode manual.

Mengutip berberapa hasil penelitian terdahulu sudah mulai menerapkan teknologi ini. Misalnya pada penelitian Fatimah dan Juanto, Melakuan analisis pemanfaatan algoritma cosine similarity sebagai sistem deteksi plagiarisme pada artikel ilmiah [1],. Penelitian serupa yang dilakukan oleh Mayola at al. Menggunakan algoritma cosine similarity untuk mendeteksi kemiripan judul disertasi yang hanya menerima dengan tingkat kemiripan 10% [9].

Penelitian Santi et al. Menggunakan pendekatan *Synonym Recognition* serupa untuk mendeteksi kemiripan tugas akhir mahasiswa menunjukan akurasi yang signifikan [10]. Meskipun pendekatan ini jarang digunakan. Namun, pengaruh *Synonym Recognition* memang sangat berdampak untuk model sistem. Tri at al. Menggunakan pendekatan berbeda dengan menggabuungkan pendekatan *Information Retrieval* unttuk deteksi plagiarisme yang dapat sistem mendeteksi kemiripan sebesar 100% [11].

Kemudian penelitian yang dilakukan oleh Sihombing dan Anggriana. Mengembangkan sistem deteksi plagiarisme dengan menggunakan pendekatan Natural Language Processing (NLP) dan Cosine Similarity dengan menetapkan treshold 0.7. Serta sistem telah berhasil mendetksi beberbeapa tugas mashasiswa [12].

Keterbatasan penelitian terdahulu terletak pada kurangnya integrasi antara representasi kontekstual teks menggunakan BERT dan pengukuran kemiripan dokumen secara matematis menggunakan cosine similarity. Kombinasi kedua pendekatan ini memungkinkan sistem untuk memahami makna kata dalam konteks kalimat, mendeteksi plagiarisme langsung maupun tidak langsung, serta diterapkan pada berbagai jenis jurnal ilmiah dengan tingkat akurasi lebih tinggi dibandingkan penelitian sebelumnya. Pendekatan BERT digunakan untuk melakukan representasi teks secara kontekstual, di mana setiap kata dalam kalimat akan dipahami berdasarkan hubungan dengan kata-kata lain di sekitarnya [13]. Hal ini memungkinkan sistem untuk menangkap makna semantik yang mendalam dan memahami parafrase atau perubahan struktur kalimat yang sering terjadi dalam praktik plagiarisme tidak langsung.

Sementara itu, Cosine Similarity berperan dalam mengukur tingkat kemiripan vektor hasil representasi teks dari BERT, sehingga dihitung seberapa besar kesamaan antar dokumen jurnal secara matematis [14]. Dengan menggabungkan kedua pendekatan tersebut, sistem diharapkan dapat melakukan deteksi plagiarisme secara akurat baik pada kata, kalimat, maupun paragraf secara lebih menyeluruh.

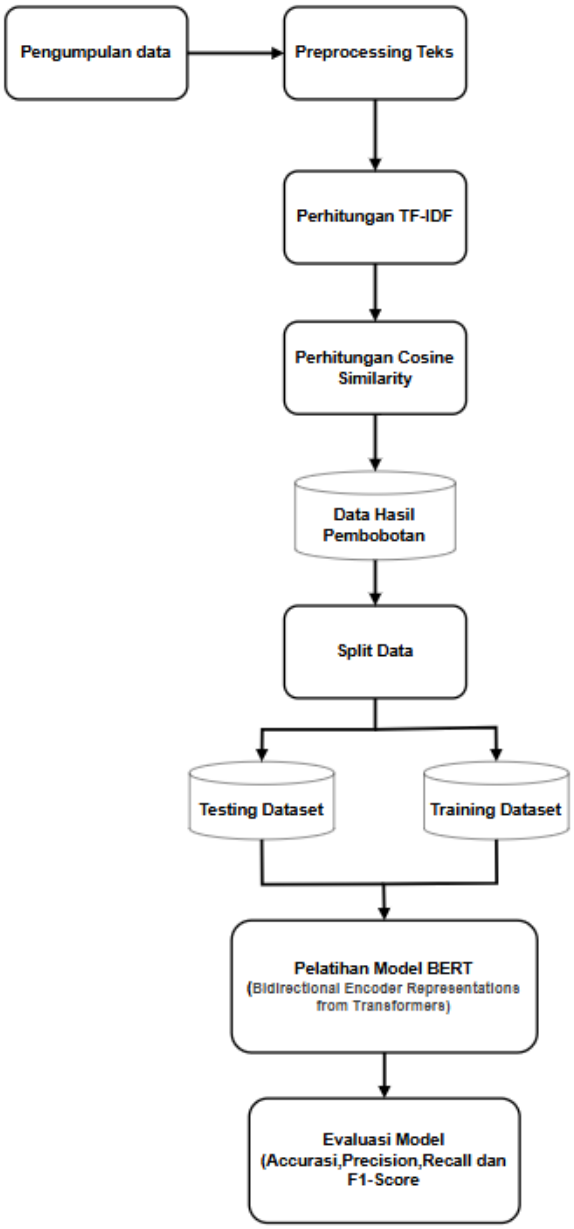
Berdasarkan keterbatasan tersebut, penelitian ini bertujuan untuk mengembangkan sistem deteksi plagiarisme dokumen jurnal berbasis BERT dan cosine similarity, sehingga proses verifikasi keaslian dokumen menjadi lebih cepat, objektif, dan andal. Sistem ini diharapkan mampu membantu institusi pendidikan dan penerbit jurnal dalam menjaga integritas karya ilmiah secara efektif.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian merupakan uraian sistematis mengenai langkah-langkah yang akan ditempuh sejak awal hingga akhir proses penelitian untuk mencapai tujuan yang telah ditetapkan[15]. Bagian ini menjadi pedoman penting dalam pelaksanaan penelitian karena menggambarkan alur kerja, pendekatan yang digunakan, serta teknik yang diterapkan dalam pengumpulan, pengolahan, dan analisis data. Dengan metode penelitian yang jelas dan terstruktur, hasil penelitian dapat dipertanggungjawabkan serta memudahkan proses replikasi atau pengembangan di masa mendatang. Berikut merupakan tahapan-tahapan yang akan dilakukan seperti pada Gambar 1.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 1. Tahapan Penelitian

Pada Gambar 1 menjelaskan tahapan yang akan dilakukan dalam penelitian ini yang terdiri dari berberapa tahap utama yang saling berhubungan, dimulai dari pengumpulan data yang berfokus untuk mengumpulkan data yang dibutuhkan. Data yang digunakan dalam penelitan ini Adalah dokumen jurnal dan artikel ilmiah yang diambil memalui google scholar dan arsip pribadi. Kemudian data yang sudah dikumpulkan akan dilakukan preprosesing seperti melakukan pembersihan data dan merubah menjadi *lowercase*. Setelah data dilakukan preprocessing Langkah selanjutnya melakukan extraksi fitur TF-IDF menjadi fektor. Yang kemudian akan dilakukan perhitungan cosine similarity. Setelah proses perhitungan cosine similarity berhasil maka akan kata atau kalimat akan dihitung jumlah kemiripannya sehingga nantinya sistem akan akan mengirim output kemiripan antara dokumen. Dan yang terakhir Adalah melakukan pelatihan model dan evaluasi model. Pada pelatihan model disini menggunakan BERT *Bidirectional Encoder Representations from Transformers.*

2.2 Pengumpuan Data

Tahap pengumpulan data merupakan langkah awal yang sangat penting dalam proses penelitian ini. Data yang digunakan dalam penelitian ini berupa dokumen teks ilmiah, khususnya jurnal akademik yang diambil dari berbagai sumber daring seperti google scolar dan repositori ilmiah. Data tersebut digunakan untuk membangun dataset yang berfungsi sebagai acuan dalam proses pelatihan dan pengujian model deteksi plagiarisme. Setelah data dokumen telah dikumpulkan maka data akan diextrak menjadi format csv.

2.2 Preprocessing Teks

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tahap preprocessing teks merupakan langkah awal yang sangat penting dalam proses deteksi plagiarisme karena menentukan kualitas data yang akan digunakan dalam analisis. Tujuannya adalah untuk mengubah teks mentah menjadi bentuk yang lebih terstruktur dan siap untuk diproses oleh algoritma. Berikut merupakan langkah-langkah preprocessing yang dilakukan antara lain:

a. *Case Folding*. Pada tahan ini yang dilakukan adalah mengubah seluruh huruf menjadi huruf kecil agar tidak ada perbedaan antara huruf besar dan kecil.
b. *Tokenization*. Pada tahap ini berfungsi untuk memecah teks menjadi potongan kata (token) sehingga dapat diproses secara individual.
c. *Stopword Removal*. Pada tahap ini bergungsi untuk menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis seperti "dan", "yang", "atau".
d. *Stemming/Lemmatization*. Pada tahap ini berfungsi untuk mengubah kata ke bentuk dasar menggunakan algoritma seperti *Sastrawi Stemmer* agar variasi kata memiliki makna yang seragam.
e. *Cleaning Text*. Pada tahap ini berfungsi untuk menghapus tanda baca, angka, dan karakter khusus yang tidak relevan.

Tahapan ini memastikan bahwa teks yang akan diolah sudah bersih dan memiliki format yang seragam untuk tahap analisis selanjutnya..

2.3 Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency merupakan metode pembobotan kata dalam dokumen yang digunakan dalam *Natural Language Processing* (NLP) [16]. Berikut merupakan formula yang digunakan untuk perhitungan *Term Frequency-Inverse Document Frequency*.

$$\text{TF(t,d)} = \frac{f(t,d)}{\sum t\prime \in f(t\prime,d)} \quad [1]$$

$$\text{IDF(t,D)} = log\frac{|D|}{1+|\{d \in D : t \in d\}|} \quad [2]$$

$$\text{TF- IDF(t,d,D)} = \text{TF(t,d)} - \text{IDF(t,D)} \quad [3]$$

Penjelasan Formula:

a. t = kata yang sedang dihitung frequensi
b. d = dokumen tertentu
c. D = jumlah total dokumen yang dikumpulkan
d. $f(t,d)$ = Jumlah kemunculan kata t dalam dokumen d
e. $\sum t' \in f(t',d)$ = total jumlah kata dalam dokumen d
f. $1 + |\{d \in D : t \in d\}|$) = jumlah dokumen yang mengandung kata t

2.4 Cosine Similarity

Cosine Similarity merupakan salah satu metode untuk mengukur tingkat kemiripan antara dua vektor dalam ruang berdimensi, dengan cara menghitung nilai kosinus dari sudut di antara kedua vektor tersebut. Metode ini tidak memperhatikan panjang vektor, melainkan hanya fokus pada arah keduanya [17]. Dalam konteks ini, semakin kecil sudut antara dua vektor, maka semakin besar nilai *cosine similarity*-nya, yang menunjukkan bahwa kedua vektor tersebut semakin mirip. Rumus dasar dari *cosine similarity* adalah hasil dari dot product antara dua vektor dibagi dengan hasil kali dari norma (panjang) masing-masing vektor. Secara matematis, cosine similarity dirumuskan sebagai berikut.

$$\text{Cos } 0 = \frac{A \cdot B}{|A| \cdot |B|} \quad [4]$$

Dimana A dan B adalah dua vektor yang dibandingkan. Nilai cosine similarity berada dalam rentang -1 hingga 1, di mana nilai 1 menunjukkan arah vektor yang identik, 0 menunjukkan kedua vektor saling tegak lurus (tidak memiliki kemiripan) [18], dan -1 menunjukkan arah yang berlawanan. Dalam praktiknya, terutama dalam pengolahan teks atau sistem rekomendasi, nilai *cosine similarity* biasanya berkisar antara 0 hingga 1 karena data umumnya bersifat non-negatif. Keunggulan utama dari *cosine similarity* adalah ketahanannya terhadap perbedaan skala data, kesederhanaannya dalam implementasi, serta efisiensinya dalam menangani data berdimensi tinggi seperti representasi dokumen dalam bentuk TF-IDF atau *bag-of-words* [19].

2.5 Split Data

Pada tahap ini, dataset dibagi menjadi dua bagian, yaitu data latih *(training set)* dan data uji (*testing set*).Tujuan dari split data ini adalah untuk memisahkan data yang digunakan untuk membangun model dan data yang digunakan untuk menguji performa model.

2.6 Pelatihan Model

Pada tahap ini, model BERT digunakan untuk melakukan embedding atau representasi teks menjadi vektor numerik yang merepresentasikan makna semantik kalimat. Proses pelatihan dilakukan dengan menyesuaikan parameter agar model mampu memahami hubungan antar kata dan konteks dalam teks jurnal.Setelah representasi vektor diperoleh, hasilnya

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

digunakan bersama *Cosine Similarity* untuk menghitung kesamaan antar dokumen.Tahapan ini bertujuan agar model mampu mengenali pola kesamaan teks, baik yang eksplisit maupun implisit (melalui parafrase atau perubahan struktur kalimat [20].

3. HASIL DAN PEMBAHASAN

Bagian ini memberikan analisis terhadap hasil eksperimen yang dilakukan dalam pengembangan sistem deteksi plagiarisme dokumen jurnal berbasis BERT dan *Cosine Similarity*. Analisis mencakup beberapa tahapan: preprocessing data, representasi teks menggunakan BERT, serta evaluasi performa sistem berdasarkan nilai Cosine Similarity. Eksperimen dilakukan menggunakan satu set data utama berupa dokumen jurnal yang telah diekstrak ke dalam format CSV, berisi pasangan teks referensi dan uji beserta label tingkat kemiripan. Setiap dokumen kemudian melalui tahap tokenisasi, embedding menggunakan BERT, dan transformasi menjadi vektor untuk penghitungan kemiripan.

3.1 Perhitungan TF-IDF

Tahapan awal dalam proses analisis kemiripan dokumen adalah menghitung *Term Frequency–Inverse Document Frequency* (TF-IDF). Tujuan dari tahap ini adalah untuk mengetahui kepentingan relatif setiap kata dalam dokumen tertentu dibandingkan seluruh dokumen di korpus, sehingga kata-kata yang sering muncul di dokumen tertentu tetapi jarang muncul di dokumen lain akan memiliki bobot lebih tinggi. Nilai TF-IDF ini menjadi dasar awal dalam representasi dokumen sebelum dilakukan embedding menggunakan BERT.

Tabel 1. Nilai TF-IDF beberapa kata kunci pada dokumen jurnal

kata	yaitu	adalah	hasil	akurasi	studi	Data	kutip
0.0	0.0	0.0	0.0	0.0	0.0	0.422	0.0
0.0	0.0	0.612	0.0	0.321	0.0	0.212	0.321
0.0	0.142	0.432	0.0	0.0	0.0	0.0	0.0
0.433	0.0	0.112	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.345	0.0	0.0	0.0	0.45
0.0	0.0	0.0	0.0	0.0	0.0	0.562	0.0
0.432	0.0	0.0	0.254	0.0	0.234	0.441	0.0
0.0	0.533	0.0	0.0	0.492	0.231	0.0	0.0
0.172	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.223	0.0	0.0	0.0	0.0	0.132	0.0	0.0
0.0	0.0	0.333	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.192	0.0

Dari Tabel 1, terlihat bahwa kata seperti data dan kutip memiliki nilai TF-IDF lebih tinggi pada beberapa dokumen, yang menunjukkan bahwa kata-kata ini memiliki bobot lebih signifikan dalam membedakan isi dokumen dibanding kata-kata umum seperti yaitu dan adalah. Nilai TF-IDF ini menjadi dasar sebelum dilakukan embedding BERT, yang selanjutnya menghasilkan representasi vektor kontekstual untuk setiap dokumen.

Tahapan selanjutnya adalah menghitung *Cosine Similarity* antar vektor dokumen untuk mengukur tingkat kemiripan. Dengan adanya representasi BERT, sistem dapat mendeteksi kemiripan tidak hanya secara literal tetapi juga makna semantik antar kalimat atau paragraf, sehingga plagiarisme langsung maupun tidak langsung dapat teridentifikasi dengan lebih akurat.

3.2 Perhitungan Cosine Similarity

Tahap berikutnya setelah perhitungan TF-IDF adalah perhitungan *Cosine Similarity*, yang bertujuan untuk mengukur tingkat kemiripan antar dokumen berdasarkan representasi vektor yang telah dihasilkan pada tahap sebelumnya. *Cosine Similarity* menghitung nilai sudut antara dua vektor dalam ruang multidimensi dan menghasilkan nilai antara 0 hingga 1, di mana:

a. Nilai mendekati 1 menunjukkan tingkat kemiripan yang sangat tinggi (dokumen cenderung identik),
b. Nilai mendekati 0 menunjukkan bahwa kedua dokumen sangat berbeda atau tidak memiliki kesamaan makna.

Dalam penelitian ini digunakan nilai *threshold* sebesar 0.6 sebagai batas pengambilan keputusan untuk menentukan apakah suatu dokumen termasuk kategori plagiat atau non-plagiat.

Tabel 2. Perhitungan Cosine Similarity

Dokumen	Cosine Similarity	Label
Dokumen 1	0.523	Non-Plagiat
Dokumen 2	0.312	Non-Plagiat
Dokumen 3	0.291	Non-Plagiat
Dokumen 4	0.333	Non-Plagiat

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9325
Hal 942-948
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Dokumen 5	0.711	Plagiat
Dokumen 6	0.894	Plagiat
Dokumen 7	0.250	Non-Plagiat

Dari Tabel 2, dapat dilihat bahwa Dokumen 5 dan Dokumen 6 memiliki nilai *Cosine Similarity* masing-masing sebesar 0.711 dan 0.894, yang melebihi ambang batas 0.6, sehingga diklasifikasikan sebagai dokumen plagiat. Sementara dokumen lainnya memiliki nilai *Cosine Similarity* di bawah 0.6 dan dikategorikan sebagai non-plagiat.

Hasil ini menunjukkan bahwa penerapan *Cosine Similarity* dengan *threshold* 0.6 efektif dalam membedakan dokumen yang memiliki kemiripan semantik tinggi dari dokumen yang berbeda secara signifikan. Dengan demikian, metode ini dapat digunakan untuk mendeteksi praktik plagiarisme baik langsung maupun tidak langsung dalam dokumen jurnal secara akurat.

3.3 Evaluasi Model BERT Tanpa TF-IDE

Pada tahap ini dilakukan evaluasi terhadap performa model BERT tanpa menggunakan metode pembobotan kata TF-IDF. Tujuannya adalah untuk mengetahui sejauh mana kemampuan model BERT murni dalam memahami konteks dan hubungan semantik antar kata tanpa adanya bantuan representasi frekuensi kata. Model BERT dievaluasi menggunakan beberapa skenario pembagian data (split data), yaitu 10:90, 20:80, dan 30:70, di mana angka pertama menunjukkan persentase data latih (*training set*) dan angka kedua menunjukkan persentase data uji (testing set). Evaluasi dilakukan menggunakan metrik Akurasi, Precision, Recall, dan F1-Score untuk menilai seberapa baik model mampu mendeteksi plagiarisme pada berbagai proporsi data. Tabel 3 menyajikan hasil evaluasi performa model BERT tanpa TF-IDF berdasarkan skenario pembagian data.

Tabel 3. Evaluasi Model Bert tanpa tf-idf

Split Data	Akurasi	Precision	Recall	F1-Score
10:90	0.8529 %	0.7126	0.6305	0.7845
20:80	0.7401	0.7012	0.5411	0.7214
30:70	0.7224	0.6522	0.5261	0.6711

Dari Tabel 3, terlihat bahwa model BERT murni mampu mendeteksi plagiarisme dengan akurasi tertinggi 85,29% pada skenario split data 10:90. Namun, akurasi dan performa model cenderung menurun ketika proporsi data latih lebih besar (20:80 dan 30:70), yang menunjukkan bahwa BERT tanpa pembobotan kata memiliki keterbatasan dalam menangkap kemiripan dokumen secara optimal.

Hasil ini mengindikasikan bahwa meskipun BERT mampu memahami konteks dan makna semantik, penerapan TF-IDF sebagai pembobot kata dapat membantu meningkatkan representasi dokumen dan performa deteksi plagiarisme secara keseluruhan, terutama pada dataset dengan variasi panjang teks dan kemiripan yang kompleks.

3.4 Evaluasi Model BERT Dengan TF-IDE

Pada tahap ini dilakukan evaluasi terhadap model BERT yang dikombinasikan dengan metode TF-IDF. Penggabungan ini bertujuan untuk meningkatkan kemampuan model dalam memahami representasi tekstual berdasarkan frekuensi kemunculan kata *(term frequency)* serta pentingnya kata dalam keseluruhan korpus (*inverse document frequency*). Dengan demikian, model tidak hanya memahami konteks semantik dari BERT, tetapi juga memperhitungkan bobot kata yang relevan dalam menentukan kemiripan antar dokumen. Hasil evaluasi model disajikan pada Tabel 4 berikut.

Tabel 4. Evaluasi Model BERT denngan tf-idf

Split Data	Akurasi	Precision	Recall	F1-Score
10:90	0.9621 %	0.8141	0.7302	0.8022
20:80	0.9501	0.9012	0.7011	0.7912
30:70	0.8820	0.9015	0.7026	0.7501

Dari Tabel 4, terlihat bahwa penggabungan BERT dengan TF-IDF menghasilkan peningkatan performa signifikan dibandingkan model BERT tanpa TF-IDF. Nilai akurasi tertinggi sebesar 96,21% diperoleh pada skenario split data 10:90, dengan Precision 0.8141, Recall 0.7302, dan F1-Score 0.8022.

Hasil ini menunjukkan bahwa model BERT dengan TF-IDF mampu mengenali pola kemiripan antar dokumen dengan tingkat ketepatan yang lebih tinggi, serta menjaga keseimbangan antara prediksi positif (plagiat) dan negatif (non-plagiat). Penggunaan TF-IDF membantu model menekankan kata-kata penting dalam dokumen sehingga meningkatkan kemampuan deteksi plagiarisme secara keseluruhan.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa sistem deteksi plagiarisme dokumen jurnal yang dikembangkan menggunakan kombinasi *Bidirectional Encoder Representations from Transformers (BERT), Term Frequency–Inverse Document Frequency (TF-IDF),* dan *Cosine Similarity* berhasil meningkatkan akurasi dan keandalan dalam membedakan

dokumen plagiat dan non-plagiat. Penerapan TF-IDF memungkinkan model menghasilkan representasi teks yang lebih bermakna sebelum diproses oleh BERT, sehingga meningkatkan kemampuan pemahaman konteks semantik. Hasil evaluasi menunjukkan bahwa kombinasi BERT dan TF-IDF dengan pembagian data 10:90 menghasilkan akurasi tertinggi 96,21%, dengan precision 0.8141, recall 0.7302, dan F1-score 0.8022, yang menunjukkan peningkatan signifikan dibandingkan penggunaan BERT tanpa TF-IDF. Selain itu, penggunaan nilai threshold 0.6 pada Cosine Similarity terbukti efektif dalam memisahkan dokumen plagiat dan non-plagiat, sementara representasi fitur yang optimal berdampak positif terhadap kecepatan komputasi dan ketepatan klasifikasi. Secara keseluruhan, penelitian ini menunjukkan bahwa integrasi BERT, TF-IDF, dan Cosine Similarity merupakan pendekatan yang lebih presisi dan efisien dibandingkan metode tradisional, sehingga sistem ini dapat dijadikan solusi komputasional yang efektif untuk mendukung deteksi plagiarisme pada jurnal akademik di era digital.

REFERENCES

[1] F. Z. Hasibuan and J. Simangunsong, "ANALISA METODE COSINE SIMILARITY DALAM MENDETEKSI PLAGIARISME PADA ARTIKEL ILMIAH," vol. 3, p. 2023.

[2] A. Tri Putra Darti Akhsa, M. Ikhwan Burhan, and A. Munandar, "Integrasi OCR dan TF-IDF untuk Metadata Otomatis pada Pencarian Dokumen Digital".

[3] F. Z. Hasibuan and J. Simangunsong, "ANALISA METODE COSINE SIMILARITY DALAM MENDETEKSI PLAGIARISME PADA ARTIKEL ILMIAH," vol. 3, p. 2023.

[4] "Musthofa Galih Pradana", doi: 10.21927/ijubi.v7i2.5170.

[5] R. Arief Permana, D. Priharsari, and A. R. Perdanakusuma, "Analisis Penggunaan Software Turnitin sebagai Alat Pendeteksi Plagiarisme," 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[6] Maulidya Prastita Syah, Ajeng Puspa Wardani, Mohammad Idhom, and Trimono, "Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks," *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 47–59, Jul. 2025, doi: 10.47709/dsi.v5i1.6021.

[7] D. Darmanto, N. I. Pradasari, and E. Wahyudi, "Sistem Deteksi Plagiarisme Tugas Akhir Mahasiswa Berbasis Natural Language Processing Menggunakan Algoritma Jaro-Winkler dan TF-IDF," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 13, no. 1, Jan. 2024, doi: 10.30591/smartcomp.v13i1.6375.

[8] E. Silalahi, D. Silalahi, D. Plagiarisme Sebagai Peningkatan, M. Irani Tarigan, and R. Veronica Sinaga, "DETEKSI PLAGIARISME SEBAGAI PENINGKATAN INTEGRITAS AKADEMIK".

[9] L. Mayola, M. Hafizh, and D. M. Putra, "Perancangan Aplikasi Similarity Deteksi Kemiripan Judul Disertasi Berbasis Web," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 2, pp. 452–257, Apr. 2024, doi: 10.47233/jteksis.v6i2.1164.

[10] S. Purwaningrum, A. Susanto, and A. Kristiningsih, "Pengaruh Synonym Recognition dalam Deteksi Kemiripan Teks Menggunakan Winnowing dan Cosine Similarity".

[11] M. Dzikry Afandi, A. Homaidi, A. Ghofur, and A. Zubairi, "Penerapan Information Retrieval dalam Sistem Analisis Kemiripan Proposal Skripsi menggunakan Cosine Similarity," *JURNAL SWABUMI*, vol. 12, no. 1, p. 2023, 2024.

[12] V. Sihombing and F. Anggriana, "Deteksi Plagiarisme Tugas Mahasiswa Menggunakan Cosine Similarity dan NLP," 2024.

[13] A. Fardhina, R. M. Siregar, M. R. W. Br Sibarani, I. C. Br Ginting, and A. Pratama, "Sistem Deteksi Berita Hoaks berbasis Algoritma Natural Language Processing (NLP) menggunakan BERT," *Jurnal Manajemen Informatika, Sistem Informasi dan Teknologi Komputer (JUMISTIK)*, vol. 4, no. 1, pp. 450–461, Jun. 2025, doi: 10.70247/jumistik.v4i1.156.

[14] A. Dzaky, H. Musta'in, A. Sanjaya, and A. B. Setiawan, "Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) 2025 154 Penerapan Regular Expression dan Cosine Similarity pada Uji Kemiripan Kalimat Bahasa Indonesia 1*," Online.

[15] I. Abdurrohim and A. Rahman, "PENERAPAN NATURAL LANGUAGE PROCESSING UNTUK ANALISIS SENTIMEN TERHADAP KEBIJAKAN PEMERINTAH".

[16] M. Hafizh Mahendra, D. Triantoro Murdiansyah, and K. Muslim Lhaksmana, "Dike : Jurnal Ilmu Multidisiplin Analisis Sentimen Tweet COVID-19 Menggunakan Metode K-Nearest Neighbors dengan Ekstraksi Fitur TF-IDF dan CountVectorizer," 2023.

[17] N. O. Idris and F. Pontoiyo, "Sistem Rekomendasi Produk Makeup Berbasis Content-Based Filtering dengan TF-IDF dan Cosine Similarity," *KETIK : Jurnal Informatika*, vol. 2, no. 06, pp. 24–32, Jul. 2025, doi: 10.70404/ketik.v2i06.311.

[18] A. Rapp, L. Di Caro, F. Meziane, and V. Sugumaran, *Natural Language Processing and Information Systems: 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25–27, 2024, Proceedings, Part I*. in Lecture Notes in Computer Science. Springer Nature Switzerland, 2024. [Online]. Available: https://books.google.co.id/books?id=mgQjEQAAQBAJ

[19] J. P. Pamput, A. R. Muthmainnah, D. F. Surianto, and N. Fadilah, "Perbandingan Cosine Similarity dan Weighted Jaccard Similarity dalam Pengembangan Mesin Pencari Perpustakaan Digital," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 10, no. 4, pp. 907–919, Sep. 2025, doi: 10.30591/jpit.v10i4.8773.

[20] O. Sanseviero, P. Cuenca, A. Passos, and J. Whitaker, *Hands-On Generative AI with Transformers and Diffusion Models*. O'Reilly Media, 2024. [Online]. Available: https://books.google.co.id/books?id=0CczEQAAQBAJ