

## PERBANDINGAN SEMANTIC SEARCH BERBASIS *SENTENCE-BERT* DAN TF-IDF PADA PENCARIAN DOKUMEN ANCAMAN SIBER

Nanda Setiawan<sup>1</sup>, Johan<sup>2</sup>, Iswanto<sup>3</sup>, Roslina<sup>4\*</sup>

<sup>1,2,3</sup>Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Potensi Utama, Medan, Indonesia

<sup>4</sup>Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Politeknik Negeri Medan, Medan, Indonesia

e-mail: <sup>1</sup>[nanda.setiawansn27@gmail.com](mailto:nanda.setiawansn27@gmail.com), <sup>2</sup>[johanst1984@gmail.com](mailto:johanst1984@gmail.com),  
<sup>3</sup>[iswantool1982@gmail.com](mailto:iswantool1982@gmail.com), <sup>4</sup>[roslina@polmed.ac.id](mailto:roslina@polmed.ac.id)

### ABSTRAK

Peningkatan jumlah dokumen ancaman siber menyebabkan proses pencarian informasi menjadi semakin kompleks karena metode berbasis kata kunci belum mampu memahami hubungan semantik antar dokumen. Penelitian ini bertujuan membandingkan performa metode *Term Frequency-Inverse Document Frequency (TF-IDF)* dan *Sentence Bidirectional Encoder Representations from Transformers (Sentence-BERT)* pada sistem pencarian dokumen ancaman siber. Tahapan penelitian meliputi impor *Dataset CyberBERT*, *preprocessing* teks, pembangunan model pencarian, pengujian menggunakan kueri, serta evaluasi hasil berdasarkan relevansi dokumen dan waktu respons. *Dataset* yang digunakan terdiri atas 1100 dokumen yang diproses menjadi representasi vektor menggunakan *TF-IDF* dan *embedding* berdimensi 384 menggunakan *Sentence-BERT*. Hasil penelitian menunjukkan bahwa kedua metode mampu mengembalikan dokumen yang relevan terhadap kueri yang diberikan, namun menghasilkan urutan dokumen yang berbeda karena perbedaan pendekatan representasi teks. Metode *TF-IDF* memiliki waktu respons lebih cepat, yaitu 0,0089 detik, sedangkan *Sentence-BERT* membutuhkan 0,0312 detik. Di sisi lain, *Sentence-BERT* lebih mampu menangkap hubungan semantik sehingga menghasilkan pencarian yang lebih kontekstual. Dengan demikian, *TF-IDF* lebih unggul dari sisi efisiensi, sedangkan *Sentence-BERT* lebih efektif dalam memahami makna dokumen sehingga sesuai untuk sistem *semantic search* pada domain ancaman siber.

**Kata kunci:** *semantic search*; *Sentence-BERT*; *TF-IDF*; pencarian dokumen; ancaman siber.

### ABSTRACT

*The increasing number of cyber threat documents has made the information retrieval process increasingly complex because keyword-based methods are unable to understand semantic relationships between documents. This study aims to compare the performance of the Term Frequency-Inverse Document Frequency (TF-IDF) and Sentence Bidirectional Encoder Representations from Transformers (Sentence-BERT) methods in a cyber threat document retrieval system. The research stages include Importing the CyberBERT Dataset, text preprocessing, building a search model, testing using queries, and evaluating the results based on document relevance and response time. The Dataset used consists of 33 documents processed into vector representations using TF-IDF and 384-dimensional embedding using Sentence-BERT. The results show that both methods are able to return relevant documents to the given query, but produce different document sequences due to differences in text representation approaches. The TF-IDF method has a faster response time of 0.0089 seconds, while Sentence-BERT requires*

0.0312 seconds. On the other hand, Sentence-BERT is better able to capture semantic relationships, resulting in more contextual searches. Thus, TF-IDF is superior in terms of efficiency, while Sentence-BERT is more effective in understanding the meaning of documents, making it suitable for semantic search systems in the cyber threat domain.

**Keywords:** semantic search; Sentence-BERT; TF-IDF; document search; cyber threats.

## 1. PENDAHULUAN

Perkembangan teknologi informasi telah meningkatkan jumlah dokumen digital yang berkaitan dengan keamanan siber, seperti laporan insiden, *Cyber Threat Intelligence (CTI)*, *vulnerability reports*, analisis *malware*, serta dokumen yang memuat *Tactics, Techniques, and Procedures (TTPs)*. Dokumen-dokumen tersebut menjadi sumber informasi penting bagi peneliti, praktisi keamanan siber, dan organisasi dalam memahami pola serangan serta menyusun strategi mitigasi. Namun, pertumbuhan jumlah dokumen menyebabkan proses pencarian informasi menjadi semakin kompleks karena pengguna tidak hanya membutuhkan dokumen yang memiliki kesamaan kata, tetapi juga kesamaan makna dan konteks.

Metode pencarian berbasis kata kunci, seperti *Term Frequency-Inverse Document Frequency (TF-IDF)*, masih memiliki keterbatasan dalam memahami hubungan semantik antar kata. *TF-IDF* hanya memberikan bobot berdasarkan frekuensi kemunculan kata pada dokumen dan korpus sehingga hasil pencarian sangat bergantung pada kesamaan istilah. Akibatnya, penggunaan sinonim, istilah teknis yang berbeda, atau variasi struktur kalimat sering menyebabkan dokumen yang sebenarnya relevan tidak berhasil ditemukan. Kondisi ini menjadi tantangan pada pencarian dokumen ancaman siber karena kosakata yang digunakan sangat beragam, dinamis, dan terus berkembang [1].

Perkembangan *Natural Language Processing (NLP)* menghadirkan pendekatan *semantic search* yang mampu memahami makna kalimat melalui representasi vektor (*embedding*). Salah satu model yang banyak digunakan adalah *Sentence Bidirectional Encoder Representations from Transformers (Sentence-BERT atau SBERT)*. Berbeda dengan *TF-IDF*, *Sentence-BERT* merepresentasikan keseluruhan kalimat ke dalam vektor berdimensi tinggi sehingga hubungan semantik antarkalimat dapat diukur menggunakan *cosine Similarity*. Dengan demikian, sistem pencarian tidak hanya mempertimbangkan kesamaan kata, tetapi juga kesamaan makna sehingga lebih mampu menemukan dokumen yang relevan meskipun menggunakan istilah yang berbeda [2].

Meningkatnya ancaman siber semakin memperkuat kebutuhan akan sistem pencarian dokumen yang lebih cerdas. Informasi ancaman berasal dari berbagai sumber, seperti *Open Source Intelligence (OSINT)*, basis data kerentanan, laporan vendor keamanan, dan publikasi ilmiah. Besarnya volume informasi menyulitkan analisis keamanan dalam menemukan informasi yang relevan secara cepat. Penelitian mengenai *EnhanceCTI* menunjukkan bahwa peningkatan jumlah data *Cyber Threat Intelligence* menyebabkan proses analisis menjadi lebih kompleks apabila masih menggunakan pendekatan pencarian konvensional sehingga diperlukan metode berbasis representasi semantik untuk meningkatkan kualitas pencarian informasi [3].

Berbagai penelitian terdahulu menunjukkan keunggulan pendekatan *semantic search*. Implementasi *Sentence-BERT* dan *Facebook AI Similarity Search (FAISS)* pada repositori akademik menghasilkan dokumen yang lebih relevan dibandingkan pencarian

berbasis kata kunci karena mampu memahami konteks kueri pengguna [4]. Penelitian lain pada pencarian dokumen PDF menggunakan *Sentence-BERT* dan *cosine Similarity* memperoleh nilai *Mean Reciprocal Rank (MRR)* sebesar 1,00, rata-rata *Precision* sebesar 0,80, dan *Recall* sebesar 0,92, lebih baik dibandingkan metode *keyword matching* yang hanya mencapai *Precision* 0,24 dan *Recall* 0,12 [5]. Selain itu, penerapan *Sentence-BERT* pada basis pengetahuan NASA meningkatkan *Mean Average Precision (MAP)* hingga 0,807 setelah dilakukan *fine-tuning*, yang menunjukkan efektivitas representasi semantik pada dokumen teknis [5].

Penelitian lain mengembangkan skema pencarian dokumen terenkripsi berbasis *Sentence-BERT* dan menjelaskan bahwa representasi *TF-IDF* menghasilkan vektor yang *sparse* serta belum mampu menangkap informasi semantik, sedangkan *Sentence-BERT* menghasilkan representasi yang lebih kaya sehingga meningkatkan kualitas pencarian berbasis makna [1]. Pada sistem verifikasi fakta, *Sentence-BERT* juga terbukti lebih efektif dibandingkan *TF-IDF* dalam mengukur kemiripan antarkalimat sehingga meningkatkan proses *evidence selection* secara otomatis [6].

Dalam domain keamanan siber, penelitian *EnhanceCTI* memanfaatkan *Sentence-BERT* untuk mengelompokkan laporan ancaman siber yang memiliki makna serupa dan memperoleh korelasi semantik sebesar 0,89, sehingga membuktikan efektivitas model tersebut dalam memahami hubungan antardokumen dari berbagai sumber [3]. Penelitian mengenai klasifikasi laporan *Cyber Threat Intelligence* berdasarkan kerangka *MITRE ATT&CK* juga menunjukkan bahwa kombinasi *TF-IDF* dengan model berbasis *BERT* mampu meningkatkan proses klasifikasi dokumen keamanan siber melalui representasi kontekstual yang lebih kaya dibandingkan pendekatan berbasis frekuensi kata [7]. Sementara itu, penelitian yang membandingkan berbagai metode ekstraksi fitur pada dokumen pola serangan siber menunjukkan bahwa setiap metode memiliki karakteristik dan keunggulan yang berbeda sesuai dengan tujuan penggunaannya. *TF-IDF* memberikan performa yang baik pada tugas tertentu, sedangkan model berbasis *transformer* lebih unggul dalam memahami konteks dan hubungan semantik antardokumen [8].

Berdasarkan penelitian-penelitian tersebut, *Sentence-BERT* telah banyak diterapkan pada repositori akademik, sistem pencarian dokumen, maupun analisis *Cyber Threat Intelligence*. Namun, masih terdapat beberapa *research gap*. Sebagian besar penelitian menggunakan *Dataset* repositori akademik, dokumen hukum, atau koleksi umum sehingga penelitian yang secara khusus membandingkan *TF-IDF* dan *Sentence-BERT* pada dokumen ancaman siber masih sangat terbatas [4]. Penelitian sebelumnya lebih berfokus pada pengembangan sistem *semantic search* dibandingkan analisis komparatif langsung antara *TF-IDF* dan *Sentence-BERT* menggunakan *Dataset* dan skenario pencarian yang sama [5]. Sebagian penelitian hanya mengevaluasi satu metode pencarian sehingga belum memberikan gambaran mengenai kelebihan dan kelemahan masing-masing metode pada dokumen ancaman siber yang kaya akan istilah teknis [3]. Evaluasi umumnya dilakukan pada domain repositori akademik atau dokumen umum sehingga hasilnya belum sepenuhnya merepresentasikan karakteristik kosakata pada domain keamanan siber [4].

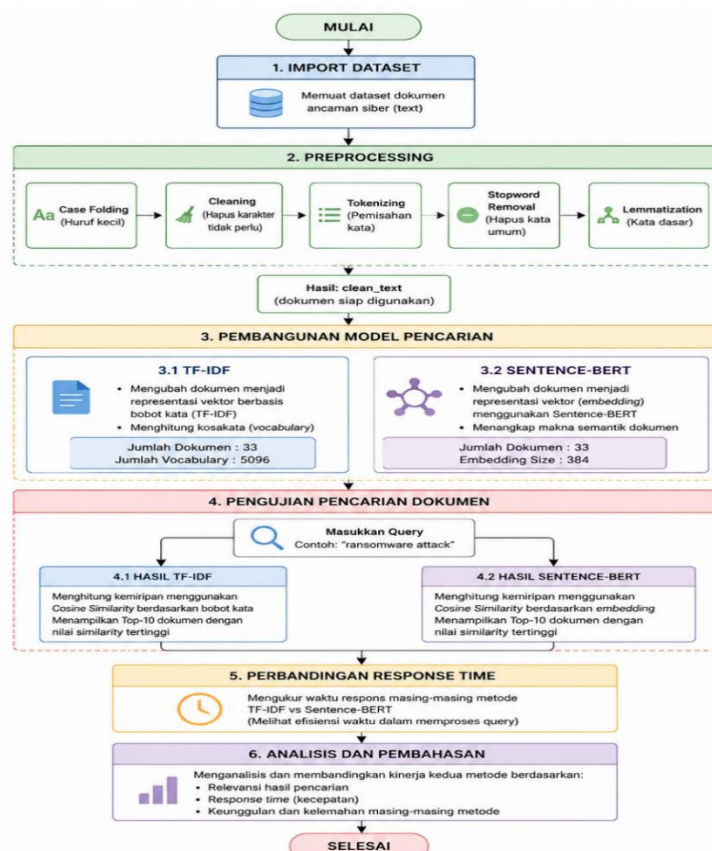
Kebaruan penelitian ini terletak pada analisis komparatif antara *TF-IDF* dan *Sentence-BERT* dalam pencarian dokumen ancaman siber menggunakan *Dataset* yang sama dan skenario pencarian yang identik. Evaluasi dilakukan secara objektif menggunakan metrik *Information Retrieval*, yaitu *Precision*, *Recall*, *Mean Reciprocal*

*Rank (MRR)*, dan *Normalized Discounted Cumulative Gain (nDCG)*, sehingga dapat memberikan gambaran yang lebih komprehensif mengenai efektivitas kedua metode.

Berdasarkan uraian tersebut, penelitian ini bertujuan membandingkan performa *TF-IDF* dan *Sentence-BERT* dalam pencarian dokumen ancaman siber. Perbandingan dilakukan untuk mengetahui kemampuan masing-masing metode dalam menghasilkan dokumen yang relevan terhadap kueri pengguna, mengevaluasi kualitas hasil pencarian menggunakan metrik *Information Retrieval*, serta memberikan rekomendasi metode yang paling efektif untuk diterapkan pada sistem temu kembali informasi pada domain keamanan siber.

## 2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini terdiri atas beberapa tahapan yang dilakukan secara sistematis untuk membangun dan mengevaluasi sistem pencarian dokumen ancaman siber menggunakan metode *TF-IDF* dan *Sentence-BERT*. Tahapan penelitian dimulai dari proses *Import Dataset*, dilanjutkan dengan *preprocessing* data, pembangunan model pencarian, pengujian menggunakan query, hingga analisis hasil pencarian dan perbandingan waktu respons dari kedua metode. Alur tahapan penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

## 1. Import Dataset dan Informasi Dataset

Tahap pertama dilakukan dengan mengimpor *Dataset CyberBERT.csv* ke dalam lingkungan Python menggunakan pustaka *Pandas*. Selanjutnya dilakukan identifikasi karakteristik *Dataset*, seperti jumlah dokumen, struktur data, tipe data, serta pemeriksaan *missing value* untuk memastikan data siap digunakan pada tahap berikutnya.

Jumlah dokumen dalam *Dataset* dihitung menggunakan Persamaan (1).

$$N = \sum_{i=1}^n 1 \quad (1)$$

dengan: (III.1)

- $N$ = jumlah dokumen dalam *Dataset*,
- $n$ = banyaknya baris data.

Penggunaan rumus: Persamaan (1) digunakan untuk mengetahui jumlah keseluruhan dokumen yang akan diproses pada penelitian. Nilai  $N$  selanjutnya digunakan pada proses pembobotan TF-IDF, khususnya dalam perhitungan *Inverse Document Frequency (IDF)*.

## 2. Preprocessing

Secara umum, proses *preprocessing* dapat dinyatakan menggunakan fungsi transformasi sebagai berikut.

$$T_{clean} = f(T) \quad (2)$$

dengan:

- $T$ = teks asli,
- $f(\cdot)$ = fungsi *preprocessing*,
- $T_{clean}$ = teks hasil *preprocessing*.

Penggunaan rumus: Persamaan (2) menunjukkan bahwa setiap dokumen diproses melalui serangkaian tahapan *preprocessing* sehingga menghasilkan teks yang lebih bersih dan konsisten. Hasil *clean text* inilah yang digunakan sebagai masukan pada proses pembentukan model TF-IDF maupun *Sentence-BERT*.

## 3. Pembangunan Model Pencarian

Pada tahap ini dibangun dua model pencarian dokumen, yaitu TF-IDF dan *Sentence-BERT*. Model TF-IDF membentuk representasi dokumen berdasarkan bobot setiap kata, sedangkan *Sentence-BERT* menghasilkan representasi dokumen dalam bentuk *embedding* yang mampu menangkap hubungan makna antar dokumen.

### a. Pembobotan TF-IDF

Pembobotan TF-IDF dihitung menggunakan Persamaan (3).

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (3)$$

(III.3)

dengan:

- $w_{t,d}$  = bobot kata  $t$  pada dokumen  $d$ ,
- $tf_{t,d}$  = frekuensi kemunculan kata,
- $N$  = jumlah dokumen,
- $df_t$  = jumlah dokumen yang mengandung kata  $t$ .

Penggunaan rumus: Persamaan (III.3) digunakan untuk menghitung bobot setiap kata pada seluruh dokumen sehingga terbentuk representasi dokumen dalam bentuk vektor TF-IDF.

b. Representasi *Sentence-BERT*

*Sentence-BERT* menghasilkan representasi dokumen menggunakan fungsi *encoder* sebagai berikut.

$$\mathbf{e} = \text{Encoder}(T_{\text{clean}}) \quad (4)$$

dengan:

- $\mathbf{e}$  = vektor *embedding*,
- $\text{Encoder}(\cdot)$  = model *Sentence-BERT*,
- $T_{\text{clean}}$  = teks hasil *preprocessing*.

Penggunaan rumus: Persamaan (III.4) digunakan untuk mengubah setiap dokumen menjadi vektor berdimensi tinggi (*embedding*) sehingga informasi semantik dokumen dapat direpresentasikan dengan lebih baik.

#### 4. Pengujian Pencarian Dokumen

Tahap pengujian dilakukan dengan memasukkan *query* pencarian ke dalam sistem. Selanjutnya, kedua model menghitung tingkat kemiripan antara *query* dan dokumen menggunakan *Cosine Similarity*, kemudian menampilkan dokumen dengan nilai *Similarity* tertinggi sebagai hasil pencarian.

Perhitungan *Cosine Similarity* dilakukan menggunakan Persamaan (5).

$$\text{CosSim}(Q, D) = \frac{Q \cdot D}{\|Q\| \|D\|} \quad (5)$$

dengan:

- $Q$  = vektor *query*,
- $D$  = vektor dokumen,
- $Q \cdot D$  = hasil *dot product*,
- $\|Q\|$  dan  $\|D\|$  = panjang vektor.

Penggunaan rumus: Persamaan (III.5) digunakan untuk menghitung tingkat kemiripan antara *query* dan setiap dokumen. Semakin tinggi nilai *Cosine Similarity*, maka dokumen dianggap semakin relevan terhadap *query* yang diberikan.

## 5. Perbandingan Response Time

Tahap ini bertujuan untuk mengukur efisiensi kedua metode berdasarkan waktu yang dibutuhkan dalam memproses *query*. Pengukuran dilakukan menggunakan Persamaan (6).

$$RT = t_{akhir} - t_{awal} \quad (6)$$

dengan:

- $RT$  = *response time*,
- $t_{awal}$  = waktu mulai proses,
- $t_{akhir}$  = waktu selesai proses.

Apabila dilakukan beberapa kali pengujian, maka rata-rata *response time* dihitung menggunakan Persamaan (7).

$$\bar{RT} = \frac{\sum_{i=1}^n RT_i}{n} \quad (7)$$

dengan:

- $\bar{RT}$  = rata-rata *response time*,
- $RT_i$  = waktu respons pada pengujian ke- $i$ ,
- $n$  = jumlah pengujian.

Penggunaan rumus: Persamaan (6) digunakan untuk memperoleh waktu respons setiap proses pencarian, sedangkan Persamaan (7) digunakan untuk menghitung rata-rata waktu respons sehingga efisiensi TF-IDF dan *Sentence-BERT* dapat dibandingkan secara objektif.

## 6. Analisis Hasil

Tahap terakhir dilakukan dengan menganalisis hasil pencarian dan membandingkan kinerja metode TF-IDF dan *Sentence-BERT* berdasarkan hasil pengujian dan pengukuran *response time*. Analisis ini digunakan sebagai dasar dalam menarik kesimpulan mengenai karakteristik dan performa masing-masing metode dalam pencarian dokumen ancaman siber.

# 3. HASIL DAN PEMBAHASAN

## Dataset Dokumen

*Dataset* yang digunakan dalam penelitian ini berupa kumpulan dokumen teks yang berkaitan dengan bidang *cybersecurity*. *Dataset* diperoleh melalui platform *Google Colaboratory (Google Colab)* sebagai lingkungan pengolahan dan pengumpulan data. Proses pengambilan *dataset* dilakukan secara terprogram menggunakan *script* pada *Google Colab*, kemudian data yang diperoleh diseleksi dan disiapkan untuk tahap pengolahan selanjutnya.

Total *dataset* yang digunakan dalam penelitian ini berjumlah **1.100 dokumen**. Dokumen-dokumen tersebut selanjutnya digunakan sebagai data penelitian dalam proses pengolahan dan pengujian sistem temu kembali informasi. Sebelum dilakukan proses pengujian, *dataset* melalui tahap prapemrosesan untuk membersihkan dan

mempersiapkan teks agar dapat diolah menggunakan metode yang digunakan dalam penelitian. *Dataset* tersebut kemudian dimanfaatkan untuk menganalisis tingkat kemiripan antar dokumen berdasarkan representasi dan karakteristik teks yang dihasilkan pada tahap pengolahan data.

## Hasil Pengolahan Dataset

**Tabel 1.** Tahapan Seleksi Dataset

| No. | Tahapan                    | Jumlah Dokumen |
|-----|----------------------------|----------------|
| 1   | Dataset awal               | 1.100          |
| 2   | Duplikasi dihapus          | 0              |
| 3   | Dokumen kosong/tidak valid | 0              |
| 4   | Dataset akhir              | 1.100          |

Berdasarkan Tabel 1, dataset awal yang digunakan dalam penelitian terdiri atas 1.100 dokumen. Tahap seleksi dilakukan untuk memastikan kelayakan data sebelum memasuki tahap *preprocessing*. Hasil seleksi menunjukkan bahwa tidak terdapat dokumen duplikat maupun dokumen kosong atau tidak valid. Dengan demikian, seluruh 1.100 dokumen (100%) dapat dipertahankan sebagai dataset akhir.

Tidak ditemukannya dokumen duplikat menunjukkan bahwa setiap dokumen memiliki karakteristik yang berbeda sehingga dapat digunakan sebagai unit informasi dalam proses *information retrieval*. Sementara itu, tidak adanya dokumen kosong atau tidak valid menunjukkan bahwa seluruh dokumen memiliki kandungan teks yang dapat diproses untuk pembentukan representasi dan perhitungan kemiripan. Kondisi tersebut juga menjaga konsistensi jumlah data pada tahap pengujian metode TF-IDF dan *Sentence-BERT*.

Penggunaan *Sentence-BERT* dalam penelitian ini memungkinkan dokumen dan *query* direpresentasikan dalam bentuk *embedding* sehingga kemiripan dapat dianalisis berdasarkan kedekatan semantik. Penelitian Fathuddin, Mandyartha, dan Nurlaili menunjukkan bahwa *Sentence-BERT* yang dikombinasikan dengan *cosine Similarity* dapat digunakan untuk pencarian dokumen berdasarkan kesamaan makna, sehingga tidak terbatas pada kecocokan kata kunci.

Temuan lain oleh [9] menunjukkan bahwa *Sentence-BERT* memiliki kemampuan yang lebih baik dalam menangkap hubungan semantik pada proses pencarian dokumen. Temuan ini sejalan dengan penelitian mengenai *information retrieval* berbasis representasi semantik yang menunjukkan bahwa model berbasis *transformer* dapat merepresentasikan informasi pada tingkat dokumen dan menangkap hubungan makna yang tidak hanya bergantung pada kecocokan kata secara langsung. Dengan menggunakan representasi berbasis *embedding*, *Sentence-BERT* memungkinkan dokumen dibandingkan berdasarkan kedekatan makna melalui representasi vektor, sehingga lebih sesuai untuk menangani variasi istilah dan hubungan semantik dalam pencarian dokumen.

Berdasarkan hasil seleksi dan penelitian terdahulu tersebut, 1.100 dokumen dinyatakan layak digunakan sebagai dataset akhir. Jumlah data yang sama pada kedua metode memungkinkan perbandingan TF-IDF dan *Sentence-BERT* dilakukan secara konsisten, sehingga perbedaan hasil pengujian dapat lebih mencerminkan karakteristik masing-masing metode.

### Hasil Representasi Dokumen Menggunakan TF-IDF

Berdasarkan Tabel 2, sistem mampu menemukan dokumen yang memiliki keterkaitan dengan setiap *query* berdasarkan nilai *Similarity*. Pada *query* “ransomware attack”, dokumen dengan peringkat pertama adalah “ransomware attack through network vulnerability” dengan nilai *Similarity* sebesar 0,66. Hasil tersebut menunjukkan adanya keterkaitan semantik yang cukup kuat karena dokumen memuat konteks serangan *ransomware* melalui kerentanan jaringan.

Pada *query* “malware detection”, dokumen “malware detected in email attachment” memperoleh nilai *Similarity* sebesar 0,53. Meskipun susunan kata pada dokumen berbeda dengan *query*, sistem tetap mampu menemukan dokumen yang berkaitan dengan konteks deteksi *malware*. Sementara itu, *query* “phishing campaign” menghasilkan dokumen “phishing email with malicious link” dengan nilai *Similarity* sebesar 0,41. Hasil ini menunjukkan bahwa sistem mampu mengidentifikasi keterkaitan antara *phishing campaign* dengan aktivitas *phishing* melalui tautan berbahaya.

Secara keseluruhan, ketiga hasil menunjukkan bahwa sistem tidak hanya mengandalkan kecocokan kata secara langsung, tetapi juga mampu menemukan dokumen berdasarkan keterkaitan konteks. Nilai *Similarity* tertinggi diperoleh pada *query* “ransomware attack” sebesar 0,66, sedangkan nilai terendah terdapat pada *query* “phishing campaign” sebesar 0,41. Perbedaan nilai tersebut menunjukkan bahwa tingkat kedekatan semantik antara *query* dan dokumen dapat berbeda bergantung pada kesesuaian konteks yang terdapat dalam masing-masing dokumen.

Temuan ini sejalan dengan penelitian Fathuddin, Mandyartha, dan Nurlaili yang menunjukkan bahwa *Sentence-BERT* dapat digunakan untuk pencarian dokumen berbasis semantik dengan memanfaatkan *embedding* dan *cosine Similarity*. Selain itu, penelitian Zuhayir, Suharso, dan Wardati menunjukkan bahwa *Sentence-BERT* memperoleh *Precision@5* sebesar 0,84, lebih tinggi dibandingkan TF-IDF sebesar 0,56 dan TF-IDF+WordNet sebesar 0,52. Kedua temuan tersebut mendukung hasil penelitian bahwa pendekatan berbasis semantik dapat membantu menemukan dokumen yang relevan meskipun terdapat perbedaan susunan atau penggunaan kata antara *query* dan dokumen.

Dengan demikian, hasil pada Tabel 2 menunjukkan bahwa metode yang digunakan mampu menghasilkan dokumen peringkat pertama yang relevan dengan *query*. Hal ini menjadi indikasi bahwa representasi semantik dapat digunakan untuk mendukung pencarian dokumen *cybersecurity* yang memiliki hubungan makna meskipun tidak selalu menggunakan kata yang identik.

**Tabel 2.** Hasil Pencarian Dokumen Berdasarkan Nilai *Similarity*

| Index | Query             | Dokumen Peringkat 1                             | <i>Similarity</i> |
|-------|-------------------|-------------------------------------------------|-------------------|
| 0     | ransomware attack | ransomware attack through network vulnerability | 0,66              |
| 1     | malware detection | malware detected in email attachment            | 0,53              |
| 2     | phishing campaign | phishing email with malicious link              | 0,41              |

### Hasil Representasi Dokumen Menggunakan *Sentence-BERT*

Berdasarkan tabel hasil pencarian, diperoleh 10 dokumen teratas dengan nilai *Similarity* antara 0,63–0,64. Dokumen dengan nilai tertinggi sebesar 0,64 terdiri atas dua dokumen berkategori *phishing*. Sementara itu, delapan dokumen lainnya memiliki nilai *Similarity* sebesar 0,63. Hasil tersebut menunjukkan bahwa seluruh dokumen yang berada pada peringkat teratas memiliki tingkat kemiripan yang relatif berdekatan.

Dari 10 dokumen yang diperoleh, terdapat 6 dokumen berkategori *phishing* dan 4 dokumen berkategori *malware*. Dominasi kategori *phishing* menunjukkan bahwa karakteristik dokumen yang memiliki kemiripan tinggi lebih banyak berkaitan dengan aktivitas *phishing*, seperti penggunaan *email*, *web*, *malicious link*, dan aktivitas kelompok ancaman. Selain kategori ancaman, dokumen juga memuat informasi mengenai sumber serangan, lokasi, media serangan, serta tindakan mitigasi seperti *quarantine*, *patch vulnerability*, dan *monitor for phishing*.

Hasil tersebut menunjukkan bahwa sistem mampu menghasilkan dokumen yang memiliki keterkaitan berdasarkan karakteristik dan konteks ancaman, bukan hanya berdasarkan satu kata tertentu. Sebagai contoh, dokumen dengan kategori *phishing* tetap memperoleh nilai *Similarity* tinggi meskipun memiliki variasi informasi seperti penggunaan *email*, *web*, atau *network*. Hal ini menunjukkan bahwa representasi semantik dapat membantu mengidentifikasi hubungan antardokumen yang memiliki konteks ancaman serupa.

Temuan ini sejalan dengan penelitian [10] yang menunjukkan bahwa *Sentence-BERT* dapat digunakan dalam pencarian dokumen berbasis semantik dengan menghasilkan representasi teks dalam bentuk *embedding* atau vektor berdimensi tetap, kemudian mengukur kedekatan antarrepresentasi menggunakan *cosine similarity*. Penelitian [11] juga menunjukkan bahwa *Sentence-BERT* dapat memetakan query dan dokumen ke dalam ruang vektor yang sama sehingga memungkinkan pengukuran kedekatan semantik untuk kebutuhan *document retrieval*. Kedua penelitian tersebut mendukung temuan bahwa pendekatan berbasis semantik mampu menangkap hubungan antarteks meskipun terdapat variasi kata dan informasi pada dokumen.

Dengan demikian, hasil 1100 dokumen teratas menunjukkan bahwa metode yang digunakan mampu mengidentifikasi dokumen *cybersecurity* yang memiliki kemiripan tinggi serta mengelompokkan hasil pencarian berdasarkan kategori ancaman yang relevan. Dominasi kategori *phishing* pada hasil pencarian juga menunjukkan bahwa karakteristik semantik dokumen tersebut memiliki kedekatan yang lebih kuat terhadap *query* yang digunakan.

**Tabel 3.** Hasil Peringkat Dokumen Berdasarkan *Similarity* dan *Threat Category*

| Ind<br>ex | Document                                                                  | <i>Simila<br/>rity</i> | <i>Threat<br/>Category</i> |
|-----------|---------------------------------------------------------------------------|------------------------|----------------------------|
| 0         | Phishing – Unknown – Email – Germany – Ransomware – Increase Web Security | 0,64                   | Phishing                   |
| 1         | Phishing – Unknown – Web – Russia – Malware – Monitor for Phishing        | 0,64                   | Phishing                   |
| 2         | Phishing – Lazarus Group – Email – Germany – Malware – Quarantine         | 0,63                   | Phishing                   |

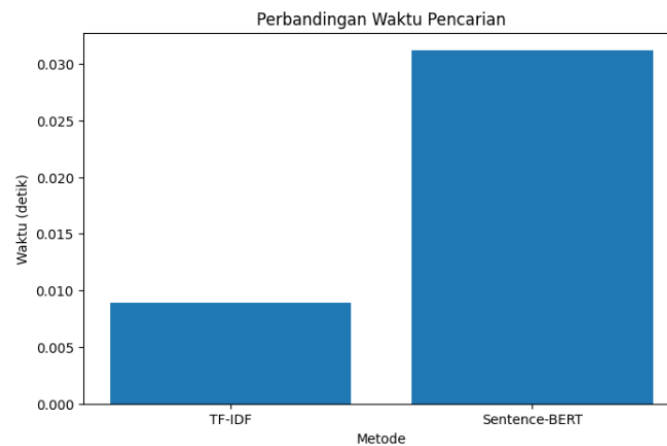
| Ind<br>ex | Document                                                                         | Simila<br>rity | Threat<br>Category |
|-----------|----------------------------------------------------------------------------------|----------------|--------------------|
| 3         | Phishing – APT-28 – Web – Russia – Malware – Patch Vulnerability                 | 0,63           | Phishing           |
| 4         | Malware – Unknown – Email – USA – DDoS – Monitor for Phishing                    | 0,63           | Malware            |
| 5         | Phishing – Lazarus Group – Email – Global – Phishing – Quarantine                | 0,63           | Phishing           |
| 6         | Malware – Unknown – Network – North Korea – DDoS – Patch Vulnerability           | 0,63           | Malware            |
| 7         | Malware – Unknown – Network – Russia – Malware – Increase Web Security           | 0,63           | Malware            |
| 8         | Phishing – Unknown – Network – USA – Ransomware – Increase Web Security          | 0,63           | Phishing           |
| 9         | Malware – Lazarus Group – Network – North Korea – Malware – Monitor for Phishing | 0,63           | Malware            |

### **Perbandingan Hasil Pencarian TF-IDF dan *Sentence-BERT***

Berdasarkan hasil pengujian pada dataset *Cybersecurity*, metode *TF-IDF* memiliki waktu pencarian sekitar 0,0089 detik, sedangkan *Sentence-BERT* membutuhkan sekitar 0,0312 detik. Hasil tersebut menunjukkan bahwa *TF-IDF* lebih cepat dengan waktu pemrosesan sekitar 3,5 kali lebih rendah dibandingkan *Sentence-BERT*. Perbedaan ini terjadi karena *TF-IDF* menggunakan representasi berbasis frekuensi kata sehingga proses pencarian lebih sederhana, sedangkan *Sentence-BERT* harus membentuk *embedding* menggunakan model Transformer. Meskipun lebih lambat, *Sentence-BERT* memiliki keunggulan dalam memahami hubungan makna antar teks sehingga lebih sesuai untuk pencarian berbasis *semantic search*. Dengan demikian, pada dataset penelitian ini *TF-IDF* lebih unggul dari sisi efisiensi waktu, sedangkan *Sentence-BERT* lebih potensial ketika kebutuhan utama adalah memahami kesamaan makna teks.

Hasil penelitian ini sejalan dengan tiga temuan penelitian sebelumnya. Pertama, [12] melalui menjelaskan bahwa *Sentence-BERT* dapat digunakan untuk menghasilkan *sentence embedding* pada tingkat kalimat, sedangkan *cosine similarity* digunakan untuk mengukur kemiripan semantik antarrepresentasi kalimat. Pendekatan tersebut memungkinkan teks direpresentasikan dalam ruang vektor sehingga kemiripan antar teks dapat dihitung berdasarkan kedekatan representasinya. Kedua, penelitian mengenai Penerapan TF-IDF dan *Cosine Similarity* pada sistem pencarian menunjukkan bahwa metode berbasis kata masih mampu menghasilkan pencarian yang relevan, khususnya pada dataset dengan jumlah dokumen yang relatif terbatas. [13] menerapkan TF-IDF dan *Cosine Similarity* pada dataset yang terdiri atas 50 dokumen dan memperoleh nilai *precision* rata-rata sebesar 83% pada pengujian terhadap beberapa query. Hasil tersebut menunjukkan bahwa kombinasi pembobotan TF-IDF dan pengukuran kemiripan *Cosine Similarity* masih efektif digunakan untuk melakukan pemeringkatan dokumen berdasarkan kesesuaian antara query dan isi dokumen. Ketiga, Penelitian mengenai *Sentence Transformers* menunjukkan bahwa model berbasis *sentence embedding* dapat menghasilkan representasi vektor yang efisien untuk menangkap informasi semantik pada teks [14]. Secara keseluruhan, temuan penelitian ini memperlihatkan bahwa pemilihan

metode perlu disesuaikan dengan tujuan sistem, yaitu *TF-IDF* untuk efisiensi pencarian dan *Sentence-BERT* ketika pemahaman semantik menjadi prioritas.



**Gambar 2.** Perbandingan Waktu Pencarian

### Analisis Waktu Komputasi

Berdasarkan Tabel 4, terdapat perbedaan waktu pencarian antara *TF-IDF* dan *Sentence-BERT*. Metode *TF-IDF* membutuhkan waktu 0,0089 detik, sedangkan *Sentence-BERT* membutuhkan waktu 0,0312 detik. Dengan demikian, *TF-IDF* memiliki waktu pencarian lebih cepat dibandingkan *Sentence-BERT*, dengan selisih waktu sebesar 0,0223 detik.

Hasil tersebut menunjukkan bahwa *TF-IDF* memiliki keunggulan dari sisi efisiensi waktu karena proses representasi dokumen menggunakan pembobotan berbasis frekuensi kata yang relatif sederhana. Sebaliknya, *Sentence-BERT* membutuhkan waktu lebih lama karena setiap dokumen dan *query* harus diubah menjadi representasi *embedding* sebelum dilakukan perhitungan kemiripan. Meskipun demikian, tambahan waktu tersebut masih berada pada rentang yang relatif kecil sehingga *Sentence-BERT* tetap dapat dipertimbangkan untuk pencarian dokumen yang membutuhkan pemahaman hubungan semantik.

Temuan ini sejalan dengan penelitian [14] yang menunjukkan bahwa *Sentence-BERT* dirancang untuk menghasilkan representasi kalimat dalam bentuk *embedding* yang dapat dimanfaatkan untuk berbagai tugas berbasis kemiripan semantik dan pencarian informasi. Namun, pembentukan representasi berbasis *transformer* memerlukan proses komputasi yang lebih besar dibandingkan metode berbasis frekuensi kata karena melibatkan proses *encoding* menggunakan model *neural network*. Sedangkan, *TF-IDF* menggunakan pembobotan berdasarkan frekuensi kemunculan suatu kata dalam dokumen dan frekuensi dokumen yang mengandung kata tersebut. Pendekatan ini relatif sederhana karena tidak memerlukan proses pembentukan representasi menggunakan model neural, sehingga dapat digunakan sebagai metode *retrieval* yang efisien ketika kecepatan pemrosesan menjadi salah satu prioritas [15].

Berdasarkan hasil tersebut, dapat disimpulkan bahwa *TF-IDF* unggul dalam efisiensi waktu pencarian, sedangkan *Sentence-BERT* menawarkan pendekatan pencarian berbasis makna dengan konsekuensi waktu pemrosesan yang lebih tinggi. Perbedaan ini

menjadi salah satu aspek penting dalam menentukan metode yang sesuai dengan kebutuhan sistem *information retrieval*.

**Tabel 4.** Perbandingan Waktu Pencarian TF-IDF dan *Sentence-BERT*

| Index | Metode               | Waktu Pencarian (detik) |
|-------|----------------------|-------------------------|
| 0     | TF-IDF               | 0,0089                  |
| 1     | <i>Sentence-BERT</i> | 0,0312                  |

#### Analisis Per Query dan Jenis Query

Analisis berdasarkan  $nDCG@10$  menunjukkan TF-IDF unggul pada 60 dari 80 query atau 75%, sedangkan *Sentence-BERT* unggul pada 20 query atau 25%. Tidak terdapat query dengan nilai yang benar-benar sama. Walaupun rata-rata TF-IDF lebih tinggi, hasil ini menunjukkan bahwa *Sentence-BERT* tetap memiliki kondisi tertentu ketika representasi semantik mampu menghasilkan ranking yang lebih baik.

**Tabel 5.** Win/Tie/Loss berdasarkan  $nDCG@10$

| Hasil                | Jumlah Query | Persentase   |
|----------------------|--------------|--------------|
| <b>Sentence-BERT</b> | <b>20</b>    | <b>25,0%</b> |
| <b>TF-IDF</b>        | <b>60</b>    | <b>75,0%</b> |
| <b>Tie</b>           | <b>0</b>     | <b>0,0%</b>  |

Sebagai contoh, Q063 pada kombinasi Geographical Location + Suggested Defense Mechanism memberikan  $nDCG@10$  sebesar 0,6653 pada *Sentence-BERT* dan 0,1799 pada TF-IDF. Pada Q094 dengan kombinasi Threat Actor + Attack Vector + Geographical Location, *Sentence-BERT* memperoleh 0,5552 sedangkan TF-IDF 0,0948. Sebaliknya, pada Q035 dengan kombinasi Attack Vector + Geographical Location, TF-IDF mencapai  $nDCG@10$  sebesar 1,0000 sedangkan *Sentence-BERT* 0,0000. Perbedaan ekstrem pada sebagian query tersebut menegaskan bahwa karakteristik query sangat memengaruhi metode yang paling efektif.

**Tabel 6.** Performa menurut struktur query

| Kriteria                                                   | N  | P@10<br>TF-IDF | P@10<br>SBERT | nDCG@10<br>TF-IDF | nDCG@10<br>SBERT |
|------------------------------------------------------------|----|----------------|---------------|-------------------|------------------|
| Attack Vector +<br>Geographical Location                   | 10 | 0,5600         | 0,2600        | 0,5701            | 0,2767           |
| Geographical Location +<br>Suggested Defense<br>Mechanism  | 11 | 0,6727         | 0,3273        | 0,6895            | 0,3430           |
| Threat Actor +<br>Geographical Location                    | 16 | 0,6438         | 0,4312        | 0,6988            | 0,4491           |
| Threat Actor + Suggested<br>Defense Mechanism              | 10 | 0,7300         | 0,4400        | 0,7658            | 0,4654           |
| Threat Actor + Attack<br>Vector + Geographical<br>Location | 31 | 0,3161         | 0,2032        | 0,3489            | 0,2105           |

| Kriteria                                    | N | P@10<br>TF-IDF | P@10<br>SBERT | nDCG@10<br>TF-IDF | nDCG@10<br>SBERT |
|---------------------------------------------|---|----------------|---------------|-------------------|------------------|
| Threat Actor + Attack Vector                | 1 | 0,6000         | 0,1000        | 0,7116            | 0,1389           |
| Attack Vector + Suggested Defense Mechanism | 1 | 0,0000         | 0,3000        | 0,0000            | 0,2497           |

TF-IDF unggul pada hampir seluruh kelompok query yang memiliki jumlah sampel lebih dari satu. Sentence-BERT unggul pada kelompok Attack Vector + Suggested Defense Mechanism, tetapi kelompok tersebut hanya terdiri atas satu query sehingga belum dapat digunakan untuk menarik generalisasi mengenai keunggulan Sentence-BERT pada tipe query tersebut.

#### 4. KESIMPULAN

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, sistem pencarian dokumen ancaman siber menggunakan metode TF-IDF dan *Sentence-BERT* berhasil dibangun dan dijalankan dengan baik. Tahapan penelitian dimulai dari proses *Import Dataset, preprocessing*, pembangunan model pencarian, pengujian menggunakan query, hingga evaluasi waktu respons. Hasil pengujian menunjukkan bahwa kedua metode mampu mengembalikan dokumen yang relevan terhadap query yang diberikan, namun menghasilkan urutan dokumen yang berbeda karena menggunakan pendekatan representasi teks yang berbeda. Metode TF-IDF menentukan tingkat kemiripan berdasarkan bobot kata, sedangkan *Sentence-BERT* memanfaatkan representasi semantik melalui *embedding* sehingga mampu memahami hubungan makna antar dokumen. Dari sisi efisiensi, TF-IDF memiliki waktu respons yang lebih cepat, yaitu 0,0089 detik, dibandingkan *Sentence-BERT* yang membutuhkan waktu 0,0312 detik. Dengan demikian, TF-IDF lebih unggul dalam kecepatan proses pencarian, sedangkan *Sentence-BERT* memberikan kemampuan representasi semantik yang lebih baik dalam memahami konteks dokumen, sehingga kedua metode memiliki kelebihan masing-masing sesuai dengan kebutuhan sistem pencarian dokumen.

#### DAFTAR PUSTAKA

- [1] Z. Hu, H. Dai, G. Yang, X. Yi, and W. Sheng, ‘Semantic-Based Multi-Keyword Ranked Search Schemes over Encrypted Cloud Data’, *Security and Communication Networks*, vol. 2022, pp. 1–15, Apr. 2022, doi: 10.1155/2022/4478618.
- [2] R. M. Holis, P. E. P. Utomo, and B. F. Hutabarat, ‘Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services’, *Brilliance*, vol. 5, no. 2, pp. 915–922, Oct. 2025, doi: 10.47709/brilliance.v5i2.7027.
- [3] S.-S. Chen, T.-W. Pai, and C.-Y. Sun, ‘EnhanceCTI: Enhanced semantic filtering and feature extraction framework for industry-specific cyber threat intelligence’, *Computers & Security*, vol. 158, p. 104649, Nov. 2025, doi: 10.1016/j.cose.2025.104649.

- [4] I. Lubis, H. Lubis, and I. Nur Wahidah, 'Implementation of Semantic Search in an Academic Repository Using Sentence-BERT and FAISS', *Sinkron*, vol. 10, no. 2, pp. 1060–1069, Apr. 2026, doi: 10.33395/sinkron.v10i2.15940.
- [5] M. A. H. Fathuddin, E. P. Mandyartha, and A. L. Nurlaili, 'Penerapan Sentence-Bert dan Cosine Similarity untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF', *RRJ*, vol. 8, no. 1, pp. 322–337, Oct. 2025, doi: 10.38035/rj.v8i1.1865.
- [6] H. Gusdevi, A. Setyanto, K. Kusriani, and E. Utami, 'Cosine Similarity-Based Evidences Selection for Fact Verification Using SBERT on the FEVER Dataset', *CogITO Smart Journal*, vol. 11, no. 1, pp. 52–66, Jun. 2025, doi: 10.31154/cogito.v11i1.917.52-66.
- [7] J. Baek, J. O, S. Jeong, and W. Kim, 'A Study on Improving the Automatic Classification Performance of Cybersecurity MITRE ATT&CK Tactics Using NLP-Based ModernBERT and BERTopic Models', *Electronics*, vol. 14, no. 22, p. 4434, Nov. 2025, doi: 10.3390/electronics14224434.
- [8] R. Othman, B. Rossi, and R. Barbara, 'A Comparison of Vulnerability Feature Extraction Methods from Textual Attack Patterns', 2024, *arXiv*. doi: 10.48550/ARXIV.2407.06753.
- [9] T. Chakraborty, V. La Gatta, V. Moscato, and G. Sperli, 'Information retrieval algorithms and neural ranking models to detect previously fact-checked information', *Neurocomputing*, vol. 557, p. 126680, Nov. 2023, doi: 10.1016/j.neucom.2023.126680.
- [10] A. L. Alfeo, M. G. C. A. Cimino, and G. Vaglini, 'Technological troubleshooting based on sentence embedding with deep transformers', *J Intell Manuf*, vol. 32, no. 6, pp. 1699–1710, Aug. 2021, doi: 10.1007/s10845-021-01797-w.
- [11] A. Esteva *et al.*, 'COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization', *npj Digit. Med.*, vol. 4, no. 1, p. 68, Apr. 2021, doi: 10.1038/s41746-021-00437-0.
- [12] C. Min, Y. Chu, L. Yang, B. Xu, and H. Lin, 'Locality Preserving Sentence Encoding', in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 3050–3060. doi: 10.18653/v1/2021.findings-emnlp.262.
- [13] R. Al Rasyid and D. H. U. Ningsih, 'Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata', *jtik*, vol. 8, no. 1, pp. 170–178, Jan. 2024, doi: 10.35870/jtik.v8i1.1416.
- [14] N. Reimers and I. Gurevych, 'Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 4512–4525. doi: 10.18653/v1/2020.emnlp-main.365.
- [15] G. H. Setiawan and I. M. B. Adnyana, 'Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models', *JAIC*, vol. 7, no. 2, pp. 252–257, Dec. 2023, doi: 10.30871/jaic.v7i2.6527.