JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika

[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

# PENERAPAN ALGORITMA TF-IDF DAN *COSINE SIMILARITY* UNTUK *QUERY* PENCARIAN SOAL MATA PELAJARAN SOSIOLOGI SMA

Kerina Putri [a,1,*], Nanda Aulia Ramadlani [b,2], Laili Cahyani [c,3]
[a,b,c] Universitas Trunojoyo Madura, Jalan Raya Telang, Kecamatan Kamal, Kabupaten Bangkalan, Jawa Timur 69162.
[1] kerinaputri4@gmail.com*; [2] nanda2210aulia@gmail.com; [3] laili.cahyani@trunojoyo.ac.id
* corresponding author

ARTICLE INFO

**Keywords**
*Information Retrieval TF-IDFCosine Similarity Question Search*

ABSTRACT

*Searching for questions in high school sociology question banks is often inefficient due to the large number of documents, causing difficulties for both students and teachers in finding relevant questions quickly. To address this problem, this study develops an Information Retrieval–based question search system using TF-IDF and Cosine Similarity to improve retrieval accuracy. The dataset consists of 350 sociology questions, which were processed through text preprocessing stages including case folding, tokenization, stopword removal, and stemming. The normalized documents were then weighted using TF-IDF and matched with user queries using Cosine Similarity to generate ranking results. System performance was evaluated using two threshold settings, namely threshold 10 and threshold 15, by measuring precision, recall, and F1-measure. The results show that threshold 10 yields higher precision but very low recall, causing many relevant documents to be missed. Meanwhile, threshold 15 achieves better balance with an average precision of 0.733, recall of 0.037, and F1-measure of 0.070, making it the most optimal configuration in this study. These findings indicate that increasing the threshold improves the system's ability to retrieve relevant documents while maintaining acceptable accuracy, and therefore threshold 15 is recommended for the sociology question retrieval system developed in this research.*

## 1. Pendahuluan

Perkembangan teknologi saat ini berkembang pesat, terutama dibidang Pendidikan[1]. Teknologi tidak hanya berperan sebagai sarana pembelajaran, tetapi juga dapat dimanfaatkan untuk menilai, mengevaluasi, serta meningkatkan kualitas hasil belajar siswa dan guru[2]. Dalam konteks tersebut, salah satu pemanfaatan teknologi yang memiliki kontribusi besar adalah pengembangan sistem pencarian soal yang mampu membantu pengguna menemukan soal secara lebih cepat, akurat, dan sesuai kebutuhan pembelajaran [3].

Pada pembelajaran Sosiologi tingkat SMA, bank soal menjadi komponen penting karena digunakan sebagai sumber latihan, evaluasi, maupun bahan persiapan penilaian.[2] Namun seiring bertambahnya jumlah soal pada setiap semester atau tahun pelajaran, guru dan siswa sering kali mengalami kesulitan dalam menemukan



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

soal yang relevan dengan topik tertentu. Hal ini terjadi karena proses pencarian masih dilakukam secara manual, yaitu dengan membuka dokumen satu per satu, membaca isi dokumen tersebut, kemudian mencocokannya dengan kebutuhan pembelajaran[3]. Proses manual ini tidak hanya memakan waktu, tetapi juga berpotensi menghasilkan pencarian yang tidak efisien serta menampilkan dokumen yang tidak relevan dengan kebutuhan pengguna[4]. Kondisi tersebut dapat menghambat proses evaluasi pembelajaran, baik dari sisi guru yang membutuhkan soal dengan cepat, maupun siswa yang memerlukan latihan sesuai kompetensi yang dipelajari[5].

Untuk mengatasi permasalahan tersebut, diperlukan sebuah sistem pencarian soal berbasis teks yang mampu melakukan pencarian secara efisien, efektif dan akurat berdasarkan *query* yang dimasukkan pengguna[6] Sistem pencarian soal yang dikembangkan dalam penelitian ini memberikan kemudahan bagi pengguna dalam menemukan soal sosiologi yang relevan, namun implementasinya tetap membutuhkan algoritma serta metode yang mampu memproses data teks secara tepat agar presisi hasil pencarian dapat tercapai[3].

Terdapat berbagai pendekatan pengelolaan data teks yang dapat digunakan dalam sistem temu kembali informasi. Namun, penelitian ini memfokuskan pada pemanfaatan metode TF-IDF sebagai teknik perhitungan bobot kata, serta *Cosine Similarity* sebagai pengukuran kemiripan antara *query* pengguna dan dokumen soal[7]. Kombinasi kedua metode tersebut diharapkan mampu menyediakan hasil pencarian yang lebih relevan, terukur, dan sesuai deng kebutuhan pencarian soal Sosiologi pada tingkat SMA[4].

Tujuan utama dari sistem pencarian ini adalah mempermudah pengguna baik guru maupun siswa dalam menemukan soal Sosiologi yang sesuai topik secara cepat dan tepat, tanpa harus menelusuri dokumen secara manual. Dengan demikian, sistem ini diharapkan dapat mendukung proses evaluasi dan kegiatan belajar melalui pencarian soal yang lebih efisien dan relevan.

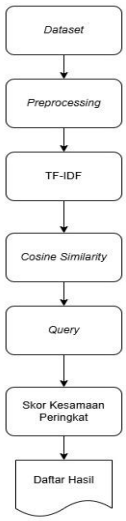
## 2. Metodologi Penelitian

Penelitian ini menggunakan pendekatan *content-based retrieval* untuk melakukan pencarian dokumen soal berdasarkan kemiripan konten antara *query* dan kumpulan dokumen soal Sosiologi SMA. Pada implementasinya, metode TF-IDF digunakan untuk memberikan bobot pada setiap *term* dalam dokumen, lalu untuk *Cosine Similarity* digunakan untuk menghitung Tingkat kemiripan antara *query* dan dan dokumen soal.[8] Hasilnya semua digunakan untuk menghasilkan ranking relevansi yang tujuannya agar soal yang paling relevan dengan *query* muncul urutan paling atas.



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]




Gambar 1. Tahapan Penelitian

### 2.1. Pengumpulan Data

Tahapan pertama dalam penelitian ini adalah proses pengumpulan dataset. Data soal diperoleh melalui metode studi pustaka dengan cara mengumpulkan berbagai soal yang telah dipublikasikan pada laman penyedia bahan ajar dan bank soal daring. Sumber utama dataset berasal dari situs https://www.websiteedukasi.com/, yang menyediakan beragam soal Sosiologi tingkat SMA untuk kelas X, XI, dan XII. Seluruh dokumen soal yang diperoleh kemudian diekstraksi ke dalam format teks agar dapat diproses lebih lanjut.

Pada penelitian ini digunakan total 350 soal, di mana setiap soal dianggap sebagai satu dokumen yang selanjutnya diproses dalam sistem Information *Retrieval* berbasis teks. Seluruh data tersebut disimpan dalam format CSV dan dikonversi menjadi dataset menggunakan *library* Python, yaitu *pandas*[9], untuk memudahkan proses pengolahan.

Selain mengumpulkan dokumen yang relevan, yaitu soal-soal Sosiologi SMA, penelitian ini juga menyertakan sejumlah dokumen tidak relevan berupa soal dari mata pelajaran Biologi. Penambahan dokumen tidak relevan ini dilakukan untuk mensimulasikan kondisi dataset yang lebih realistis dan untuk menguji kemampuan sistem dalam membedakan dokumen relevan dan tidak relevan selama proses pencarian.

### 2.2. *Preprocessing*

Tahap ini bertujuan untuk menormalisasikan setiap dokumen dapat diubah menjadi representasi numerik untuk diproses lebih lanjut. Dataset penelitian terdiri 350 soal Sosiologi SMA, dimana setiap soal dijadikan satu dokumen. Adapun tahapan *preprocessing* yang diterapkan adalah sebagai berikut:

1. *Case Folding*
   Teks diubah ke huruf kecil dan dibersihkan dari angka, simbol, maupun karakter non-alfabet menggunakan *Regular Expression* (Regex). Langkah ini bertujuan menyederhanakan bentuk teks agar fokus hanya pada kata yang relevan.[10]
2. *Tokenizing dan Filtering*



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Dokumen yang telah dibersihkan diubah menjadi token kata menggunakan metode pemisahan sederhana. Token yang terlalu pendek, tidak alfabet, atau terdeteksi sebagai *typo* dibuang agar hanya menyisakan kata yang bermakna.[11]

3. *Stopword Removal*
   Kata-kata umum yang sering muncul namun tidak memiliki kontribusi penting baik dari daftar stopword NLTK maupun tambahan manual dihilangkan. Hal ini dilakukan untuk meningkatkan kualitas fitur yang dianalisis oleh sistem.[8]
4. *Stemming*
   Setiap token kemudian diubah ke bentuk dasarnya menggunakan stemmer Sastrawi. Proses ini berguna untuk menyatukan berbagai variasi kata berimbuhan sehingga representasi dokumen menjadi lebih konsisten[9]

### 2.3. TF-IDF

Metode ini digunakan untuk menentukan bobot suatu istilah terkait yang digunakan dengan sebuah dokumen pendekatan ini menggabungkan dua prinsip untuk menghitung bobot, yaitu frekuensi kemunculan suatu istilah dalam dokumen tertentu dan frekuensi invers dari dokumen yang mengandung istilah tersebut[12]. Mengukur frekuensi kemunculan kata dalam suatu dokumen memberikan wawasan tentang signifikansi istilah dalam dokumen tersebut[9]. TF-IDF adalah istilah gabungan yang terdiri dari dua kata yang berbeda: *Term Frequency* dan *Inverse Document Frequency*[8].

TF digunakan untuk mengukur frekuensi kemunculan suatu istilah dalam dokumen, dengan menggunakan rumus dibawah ini:[9]

$$TF = \frac{jumlah\ kemunculan\ kata\ dalam\ dokumen}{jumlah\ kata\ dalam\ dokumen} \quad (1)$$

Inverse document frequency (IDF) memberikan bobot yang lebih rendah untuk kata-kata yang sering muncul dan memberikan bobot yang lebih tinggi untuk kata-kata yang jarang muncul. IDF atau inverse document frequency, berperan[6]. IDF memberikan bobot untuk kata-kata berdasarkan frekuensi kemunculannya dalam satu set dokumen[7]. Kata-kata yang sering muncul diberi bobot yang lebih rendah sementara kata-kata yang jarang muncul diberi bobot yang lebih tinggi. Berikut rumusnya :[8]

$$IDF = log\ \frac{\text{Total jumlah dokumen}}{\text{Jumlah dokumen di mana kata } t \text{ muncul}} \quad (2)$$

Bobot akhir dihitung dengan mengalikan nilai TF dan IDF: [9] (3)

$$tf - idf(t,d) = tf \times idf$$

### 2.4. *Cosine Similarity*

Skor *similarity* diperoleh dengan mengukur skor similarity antara dua vektor, yaitu vektor queri dan vektor dokumen. Semakin besar nilai relevansi, semakin mirip atau relevan queri dan dokumen tersebut[7]. Ukuran *similarity* digunakan untuk menentukan tingkat similarity antara titik data *Cosine Similarity*


https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

ditentukan sebagai kosinus sudut θ yang terbentuk antara vektor-vektor[8]. Formula *Cosine Similarity* memiliki rumus sebagai berikut :[11]

$$\text{Sim}(Q,D) = \frac{Q \cdot D}{|Q|\,|D|} = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} Q_i^2}\sqrt{\sum_{i=1}^{n} D_i^2}} \quad (4)$$

Keterangan:

- Q = Vektor *Query* (kata kunci pencarian).
- D = Vektor Dokumen (soal).
- Q · D = *Dot product* antara vektor *query* dan dokumen.
- ||Q|| dan ||D|| = *Euclidean length* (magnitudo) dari vektor.

### 2.5. Evaluasi

Evaluasi dilakukan untuk mengukur tingkat keefektifan sistem dalam menampilkan hasil pencarian soal Sosiologi yang relevan dan tidak relevan berdasarkan nilai *similarity*[7]. Evaluasi pada penilitian ini membutuhkan sebuah matriks yang disebut berupa matriks confusion[12]. Matriks *confusion* ditunjukkan pada tabel 1:

Tabel 1: *Matriks Confusion*

| Ck | *Classifier positive label* | *Classifier negative label* |
|---|---|---|
| *True positive label* | A | B |
| *True negative label* | C | D |

Keterangan:

- A (*True Positive*): jumlah dokumen yang berhasil dikategorikan oleh sistem ke dalam kategori Ck.
- B (*False Negative*): jumlah dokumen yang mempunyai kategori Ck, namun sistem tidak mengklasifikasikannya ke dalam kategori Ck.
- C (*False Positive*): jumlah dokumen bukan kategori Ck, namun sistem mengklasifikasikannya ke dalam kategori Ck.
- D (*True Negative*): jumlah dokumen yang tidak termasuk kategori Ck, dan sistem juga tidak mengklasifikasikannya ke dalam kategori Ck.

Confusion matrix ini digunakan untuk menghitung nilai *Precision*, *Recall*, dan *F-Measure*:

*Precision* adalah proporsi jumlah dokumen yang ditemukan dan dianggap relevan untuk kebutuhan si pencari informasi[7]. Rumusnya:[9]

$$Precision = \frac{A}{A+C} \quad (5)$$

Keterangan:

- A (*True Positive*)
  Jumlah dokumen yang benar-benar termasuk kategori Ck dan diklasifikasikan dengan benar oleh sistem.
- C (*False Positive*)


https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Jumlah dokumen yang bukan termasuk kategori Ck, tetapi secara salah diklasifikasikan oleh sistem sebagai kategori Ck.

- A + C

  Total semua dokumen yang diprediksi sistem sebagai kategori Ck, baik yang benar maupun yang salah.

*Recall* adalah proporsi jumlah dokumen yang dapat ditemukan-kembali oleh sebuah proses pencarian di sistem IR[9]. Adapun perhitungan *recall*: [8]

$$Recall = \frac{A}{A+B} \quad (6)$$

Keterangan:

- A (*True Positive*)
  Jumlah dokumen yang benar-benar termasuk kategori Ck dan berhasil diklasifikasikan dengan benar oleh sistem.
- B (*False Negative*)
  Jumlah dokumen yang seharusnya termasuk kategori Ck, tetapi tidak diklasifikasikan oleh sistem ke dalam kategori Ck.
- A + B
  Total semua dokumen yang aslinya memang kategori Ck, baik yang terdeteksi maupun yang tidak terdeteksi oleh sistem.

*F-Measure* adalah ukuran gabungan yang mempertimbangkan *Precision* dan *Recall* sekaligus, untuk memberikan gambaran performa sistem secara keseluruhan[12]. *F-Measure* mengharmonisasi kedua metrik ini sehingga sistem yang memiliki *Precision* tinggi tapi *Recall* rendah atau sebaliknya, tetap dapat dinilai secara seimbang.[6]

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Keterangan :

- *Precision*: seberapa tepat prediksi sistem untuk kategori Ck.
- *Recall*: seberapa lengkap sistem mengenali dokumen sebenarnya dalam kategori Ck.
- *F-Measure*: nilai akhir yang menggabungkan ketepatan dan kelengkapan prediksi. Nilainya berada di antara 0 dan 1; semakin mendekati 1, performa sistem semakin baik.

## 3. Hasil dan Pembahasan

### 3.1. Hasil Pengumpulan Data

Pada tahap pengumpulan data, seluruh soal yang diambil dari situs *websiteedukasi.com* berhasil dimuat ke dalam sistem menggunakan library pandas. Dataset terdiri dari 350 dokumen, masing-masing merepresentasikan satu butir soal yang digunakan sebagai bahan uji pada sistem pencarian. Struktur dataset memuat informasi dasar seperti nomor soal, kelas, mata pelajaran, dan isi teks soal.



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Selain soal Sosiologi sebagai dokumen relevan, dataset juga mencakup beberapa soal Biologi sebagai dokumen tidak relevan. Penyertaan dua jenis dokumen ini bertujuan untuk menguji kemampuan sistem dalam membedakan konten yang sesuai dan tidak sesuai dengan topik pencarian. Tabel 2 menampilkan contoh data yang digunakan dalam penelitian ini.

Tabel 2. *Dataset* Soal Sosiologi

| nomor | kelas | mapel | soal |
|---|---|---|---|
| 1 | 12 | Sosiologi | Salah satu faktor pendorong terjadinya perubahan sosial yang berasal dari luar masyarakat adalah .... |
| 2 | 12 | Sosiologi | Krisis minyak bumi mendorong masyarakat Indonesia melakukan inovasi/penemuan baru. Masyarakat berhasil menemukan minyak gas yang bahan bakunya dari kotoran hewan dan dapat dimanfaatkan untuk menganti minyak tanah. Contoh tersebut merupakan perubahan sosial yang disebabkan oleh faktor .... |
| 3 | 12 | Sosiologi | Berubahnya sistem pemerintahan dari sistem kerajaan menjadi presidensiil, termasuk bentuk perubahan sosial .... |
| ... | ... | ... | ... |
| 349 | 12 | Biologi | Bagaimana hubungan antara perilaku sosial dan keberhasilan suatu populasi? |
| 350 | 12 | Biologi | Jelaskan bagaimana mekanisme biologis mengatur interaksi dalam kelompok organisme. |

### 3.2. Hasil *Preprocessing*

Dataset yang digunakan dalam penelitian ini terdiri dari 350 soal Sosiologi SMA. Setiap dokumen merepresentasikan satu butir soal. Sebelum masuk ke tahap perhitungan TF-IDF dan Cosine Similarity, seluruh dokumen diproses menggunakan tahapan preprocessing agar teks menjadi lebih bersih dan seragam. Tahapan yang dilakukan meliputi: Case Folding, *Tokenizing dan Filtering,* Stopword Removal, dan Stemming.

Implementasi kode preprocessing ditunjukkan pada potongan program berikut;



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

```
!pip install Sastrawi

import pandas as pd
import numpy as np
import re
import nltk

from nltk.corpus import stopwords
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

# Case Folding
def case_folding(text):
    text = str(text).lower()
    text = re.sub(r'[^a-z\s]', ' ', text)
    text = re.sub(r'\s+', ' ', text)
    return text.strip()

df['Case_Folding'] = df['soal'].astype(str).apply(case_folding)

# Tokenizing
def tokenizing(text):
    return text.split()
df['Tokenizing'] = df['Case_Folding'].apply(tokenizing)

# Filtering
def filtering(tokens):
    hasil = []
    for t in tokens:
        if t.isalpha() and len(t) >= 3:
            hasil.append(t)
              return hasil
```



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

```
    df["Filtering"] = df["Tokenizing"].apply(filtering)

# Hapus typo absurd
typo_buruk = {"xv","aaa","bbb","ccc","zzz","qqq","lll","mmm","nnn","yt","ppp"}
df["Filtering"] = df["Filtering"].apply(lambda x: [t for t in x if t not in
typo_buruk])

# Stopword Removal
nltk.download('stopwords')
stop_words = set(stopwords.words("indonesian"))
stop_tambahan = {
    "apa","yang","pada","dalam","untuk","dengan","agar","jika","karena",
    "bagaimana","sebutkan","jelaskan","manakah","berikut","adalah"
}
stop_words |= stop_tambahan

def stopword_removal(tokens):
    return [t for t in tokens if t not in stop_words]

df["Stopword_Removal"] = df["Filtering"].apply(stopword_removal)

# Stemming
factory = StemmerFactory()
stemmer = factory.create_stemmer()

def stemming(tokens):
    return [stemmer.stem(t) for t in tokens]

df["Stemming"] = df["Stopword_Removal"].apply(stemming)

# Gabungkan kembali
def gabung_teks(tokens):
    return " ".join(tokens)

df["Preprocessed_Text"] = df["Stemming"].apply(gabung_teks)
```

Kode Sumber 1. *Preprocessing*

Kemudian hasil *preprocessing* dataset dapat dilihat pada Gambar 2:

| | Nomor | kelas | Mapel | soal | Preprocessed_Text |
|---|---|---|---|---|---|
| **0** | 1 | 12 | Sosiologi | Salah satu faktor pendorong terjadinya perubah... | salah faktor dorong ubah sosial asal masyarakat |
| **1** | 2 | 12 | Sosiologi | Krisis minyak bumi mendorong masyarakat Indone... | krisis minyak bumi dorong masyarakat indonesia... |
| **2** | 3 | 12 | Sosiologi | Berubahnya sistem pemerintahan dari sistem ker... | ubah sistem perintah sistem raja presidensiil ... |
| **3** | 4 | 12 | Sosiologi | Unsur budaya asing yang masuk tanpa seleksi da... | unsur budaya asing masuk seleksi sebab luntur ... |
| **4** | 5 | 12 | Sosiologi | Berikut ini yang tidak termasuk contoh perubah... | contoh ubah sosial |

Gambar 2. Hasil *Preprocessing*



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

### 3.3. Implementasi Pembobotan TF-IDF

Setelah proses preprocessing menghasilkan teks soal yang bersih dan seragam, tahap selanjutnya adalah melakukan pembobotan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Metode ini mengubah setiap dokumen menjadi representasi vektor numerik berdasarkan frekuensi kata dan tingkat kepentingan kata dalam keseluruhan koleksi dokumen. Perhitungan TF-IDF dilakukan menggunakan TfidfVectorizer dari library scikit-learn, dengan input berupa teks hasil preprocessing pada kolom clean_text. Hasil proses ini berupa matriks TF-IDF yang merepresentasikan bobot setiap term pada seluruh dokumen, yang kemudian digunakan sebagai dasar dalam perhitungan Cosine Similarity. Potongan kode implementasi TF-IDF dapat dilihat pada program berikut :

```
from sklearn.feature_extraction.text import TfidfVectorizer

print("[1/3] Menghitung bobot TF-IDF...")

vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(df["Preprocessed_Text"])

feature_names = vectorizer.get_feature_names_out()

tfidf_df = pd.DataFrame(
    tfidf_matrix.toarray(),
    columns=feature_names
)

print("TF-IDF Selesai! Matriks ukuran:", tfidf_matrix.shape)
```

Kode Sumber 2. TF-IDF

Hasil pembobotan TF-IDF menghasilkan representasi vektor dari setiap dokumen soal, di mana setiap nilai menunjukkan tingkat kepentingan suatu term dalam dokumen tersebut. Representasi vektor ini selanjutnya digunakan sebagai dasar perhitungan kemiripan dan proses pemeringkatan soal pada tahap berikutnya. Contoh hasil perhitungan Cosine Similarity ditampilkan pada Gambar 3.

| | abstraksi | acara | ada | adab | adan | adaptasi | adat | adenine | adi | adil |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 345 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 346 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 347 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 348 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 349 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Gambar 3. Hasil TF-IDF



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

### 3.4. Implimentasi *Cosine Similarity*

Cosine Similarity digunakan untuk mengukur kedekatan dua vektor TF-IDF berdasarkan sudut antar vektornya. Nilai similarity berkisar dari 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa dua soal memiliki kemiripan tinggi. Pada penelitian ini, matriks Cosine Similarity dihitung menggunakan fungsi cosine_similarity dari Scikit-Learn dan digunakan untuk mengidentifikasi soal paling relevan terhadap sebuah query. Implementasi kode ditunjukkan pada potongan program berikut;

```
from sklearn.metrics.pairwise import cosine_similarity

print("[2/3] Menghitung Cosine Similarity...")

cosine_sim_matrix = cosine_similarity(tfidf_matrix)

labels = [f"Soal_{i+1}" for i in range(len(df))]

cosine_df = pd.DataFrame(
    cosine_sim_matrix,
    index=labels,
    columns=labels
)

print("Cosine similarity selesai! Matriks ukuran:", cosine_sim_matrix.shape)
```

Kode Sumber 3. *Cosine Similarity*

Hasil perhitungan ini membentuk sebuah matriks kemiripan antar dokumen, yang kemudian digunakan untuk menentukan urutan soal yang paling relevan berdasarkan nilai similarity tertinggi. Contoh hasil perhitungan Cosine Similarity ditampilkan pada Gambar 4.

| | Soal_1 | Soal_2 | Soal_3 | Soal_4 | Soal_5 | Soal_6 | Soal_7 | Soal_8 | Soal_9 | Soal_10 | ... | Soal_341 | Soal_342 | Soal_343 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Soal_1** | 1.000000 | 0.193004 | 0.139412 | 0.074467 | 0.271784 | 0.045565 | 0.038660 | 0.0 | 0.129047 | 0.062511 | ... | 0.000000 | 0.0 | 0.000000 |
| **Soal_2** | 0.193004 | 1.000000 | 0.043165 | 0.046133 | 0.165110 | 0.014108 | 0.023940 | 0.0 | 0.039956 | 0.019355 | ... | 0.000000 | 0.0 | 0.000000 |
| **Soal_3** | 0.139412 | 0.043165 | 1.000000 | 0.102735 | 0.261434 | 0.043830 | 0.000000 | 0.0 | 0.000000 | 0.060131 | ... | 0.051572 | 0.0 | 0.045440 |
| **Soal_4** | 0.074467 | 0.046133 | 0.102735 | 1.000000 | 0.139646 | 0.023412 | 0.363512 | 0.0 | 0.000000 | 0.032119 | ... | 0.045576 | 0.0 | 0.040157 |
| **Soal_5** | 0.271784 | 0.165110 | 0.261434 | 0.139646 | 1.000000 | 0.085447 | 0.000000 | 0.0 | 0.000000 | 0.117226 | ... | 0.000000 | 0.0 | 0.000000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **Soal_346** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.000000 |
| **Soal_347** | 0.000000 | 0.000000 | 0.042174 | 0.037271 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | ... | 0.061796 | 0.0 | 0.054449 |
| **Soal_348** | 0.000000 | 0.016695 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.000000 |
| **Soal_349** | 0.000000 | 0.016181 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.000000 |
| **Soal_350** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.000000 |

Gambar 4. Hasil *Cosine Similarity*

### 3.5. Hasil Pengujian Kinerja Sistem

*Query* yang digunakan dalam pengujian ini adalah sebagai berikut:



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Tabel 3. *Query*

| query |
|---|
| interaksi sosial dan faktor yang mempengaruhi |
| nilai dan norma sosial dalam kehidupan masyarakat |
| perilaku menyimpang dan pengendalian social |
| kelompok sosial dan karakteristiknya |
| mobilitas sosial di masyarakat modern |
| perubahan sosial dan dampaknya |
| lembaga sosial dan perannya |
| masyarakat multikultural Indonesia |
| sosialisasi dan pembentukan kepribadian |
| hubungan masyarakat dengan lingkungannya |

### 3.5.1 Hasil Pengujian *Query*

Hasil pencarian dokumen dengan *query* interaksi sosial dan faktor yang mempengaruhi

Table 4. Hasil Pencarian Query interaksi sosial dan faktor yang mempengaruhi

| Threshold | Kelas | Mapel | soal | Relevansi |
|---|---|---|---|---|
| 10 | 12 | Sosiologi | Terjadinya mobilitas sosial biasanya dipengaruhi oleh faktor .... | Relevan |
| | 12 | Biologi | Bagaimana pencemaran lingkungan memengaruhi interaksi antarorganisme? | Tidak Relevan |
| | 11 | Sosiologi | Konflik pada dasarnya merupakan suatu interaksi sosial yang bersifat... | Relevan |
| | 10 | Sosiologi | Suatu proses sosial atau interaksi sosial disebut asosiatif, jika.... | Relevan |
| | 12 | Sosiologi | Bagaimana interaksi antarorganisme memengaruhi keseimbangan suatu ekosistem? | Tidak Relevan |
| | 11 | Biologi | Berikut yang merupakan faktor-faktor yang dapat mempercepat terjadinya integrasi sosial... | Relevan |
| | 12 | Sosiologi | Bagaimana kondisi lingkungan memengaruhi pola interaksi hewan di habitatnya? | Tidak Relevan |
| | 12 | Biologi | Di bawah ini merupakan faktor yang mempengaruhi terjadinya difusi intramasyarakat adalah .... | Relevan |
| | 12 | Sosiologi | Berikut ini yang bukan faktor yang memengaruhi difusi antarmasyarakat adalah... | Relevan |
| | 10 | Sosiologi | Syarat utama terjadinya interaksi sosial adalah adanya kontak sosial dan... | Relevan |
| 15 | 12 | Sosiologi | Berikut ini merupakan faktor-faktor yang dapat menyebabkan perubahan sosial budaya, kecuai .... | Relevan |



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

| | | | |
|---|---|---|---|
| 12 | Biologi | Faktor apa yang memengaruhi pembentukan hierarki dalam kelompok hewan? | Tidak Relevan |
| 12 | Sosiologi | Unsur yang terdapat dalam masyarakat yang mampu mempengaruhi perubahan sosial disebut dengan ... | Relevan |
| 10 | Sosiologi | Berikut ini merupakan bentuk interaksi sosial yang bersifat disosiatif, kecuali ..... | Relevan |
| 12 | Biologi | Apa faktor biologis yang memengaruhi pembentukan pemimpin dalam kelompok hewan? | Tidak Relevan |

Table 5. Hasil Pengujian dengan Threshold 10

| | ***Retrieved*** | ***Non-Retrieved*** |
|---|---|---|
| ***Actual*** | 7 | 293 |
| ***Non-Actual*** | 3 | 47 |

*Precision* = *0,*700
*Recal* = *0,*023
*F-measure* = 0,045

Table 6. Hasil Pengujian dengan Threshold 15

| | ***Retrieved*** | ***Non-Retrieved*** |
|---|---|---|
| ***Actual*** | 10 | 290 |
| ***Non-Actual*** | 5 | 45 |

*Precision* = 0,667
*Recal* = 0,033
F-*measure* = 0,063

Kemudian *Query* 2 sampai *Query* 10 dilakukan perhitungan seperti pada *Query* 1 di atas untuk mencari nilai precision, *recall* dan *f.measure* pada threshold 10 dan 15. Hasil dari pencarian precision, *recall* dan *f1-measure* pada semua *query* dapat dilihat pada tabel berikut:

Table 7. Hasil Evaluasi *Precision, Recall, F-measure* pada *Threshold* 10

| | ***Precision*** | ***Recall*** | ***F-measure*** |
|---|---|---|---|
| *Query 1* | 0,700 | 0,023 | 0,045 |
| *Query 2* | 1,000 | 0,033 | 0,065 |
| *Query 3* | 0,200 | 0,007 | 0,013 |
| *Query 4* | 0,900 | 0,030 | 0,058 |
| *Query 5* | 1,000 | 0,033 | 0,065 |
| *Query 6* | 1,000 | 0,033 | 0,065 |
| *Query 7* | 0,800 | 0,027 | 0,052 |
| *Query 8* | 1,000 | 0,033 | 0,065 |
| *Query 9* | 0,900 | 0.03 | 0,058 |
| *Query 10* | 0,200 | 0,007 | 0,013 |
| Rata-rata | **0,770** | **0,025** | **0,050** |

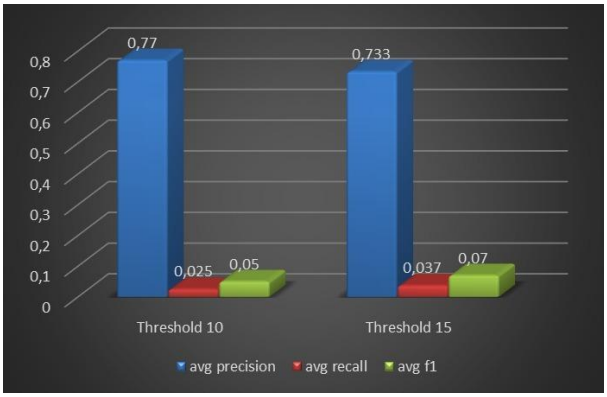


https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

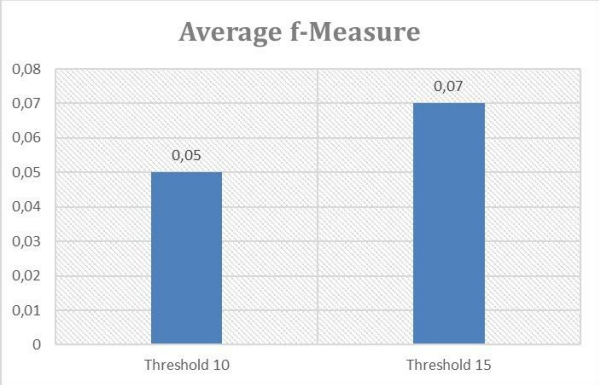
JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 20 No 1]
[Januari 2026]

Table 8. Hasil Evaluasi *Precision, Recall, F-measure* pada *Threshold* 15

| | ***Precision*** | ***Recall*** | ***F-measure*** |
|---|---|---|---|
| *Query 1* | 0,667 | 0,033 | 0,063 |
| *Query 2* | 0,933 | 0,047 | 0,089 |
| *Query 3* | 0,200 | 0,010 | 0,019 |
| *Query 4* | 0,733 | 0,037 | 0,070 |
| *Query 5* | 1,000 | 0,050 | 0,095 |
| *Query 6* | 1,000 | 0,050 | 0,095 |
| *Query 7* | 0,733 | 0,037 | 0,070 |
| *Query 8* | 1,000 | 0,050 | 0,095 |
| *Query 9* | 0,867 | 0,043 | 0,083 |
| *Query 10* | 0,200 | 0,010 | 0,019 |
| Rata-rata | **0,733** | **0,037** | **0,070** |




Gambar 5. Grafik *average* dari *precision, recall* dan *f.mesure* pada setiap *threshold*




Gambar 6. Grafik nilai *f.measure* dari *threshold* 10 dan 15

### 3.6. Hasil Analisis Pengujian

Berdasarkan hasil pengujian yang dilakukan pada dua threshold, yaitu threshold 10 dan threshold 15, terlihat adanya perbedaan performa sistem dalam melakukan temu kembali dokumen. Pada threshold 10, nilai *precision* berada pada angka rata-rata 0.770, sedangkan pada threshold 15 nilai *precision* sedikit menurun



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

menjadi 0.733. Penurunan *precision* ini wajar terjadi karena semakin tinggi threshold, semakin banyak dokumen yang diambil, sehingga peluang masuknya dokumen yang tidak relevan menjadi lebih besar.

Berbeda dengan precision, nilai *recall* justru mengalami peningkatan ketika threshold dinaikkan. Threshold 10 hanya menghasilkan *recall* sebesar 0.025, sedangkan threshold 15 meningkat menjadi 0.037. Kondisi ini menunjukkan bahwa dengan mengambil lebih banyak dokumen, sistem memiliki peluang lebih besar untuk menemukan dokumen relevan yang sebelumnya terlewat pada threshold yang lebih rendah.

Kombinasi perubahan *precision* dan *recall* ini juga berdampak pada nilai F1-measure. Pada threshold 10, *F1-measure* tercatat sebesar 0.050, sementara threshold 15 menghasilkan nilai F1-measure tertinggi yaitu 0.070. Meskipun *precision* menurun, peningkatan *recall* yang cukup signifikan membuat threshold 15 menghasilkan keseimbangan performa yang lebih baik. Dengan demikian, secara keseluruhan threshold 15 dapat dinilai sebagai konfigurasi yang lebih optimal karena mampu mempertahankan tingkat *precision* yang cukup baik sambil meningkatkan kemampuan sistem dalam menemukan dokumen relevan.

## 4. Kesimpulan

Berdasarkan keseluruhan proses penelitian dan pengujian yang telah dilakukan menggunakan metode TF-IDF dan *cosine similarity*, dapat disimpulkan bahwa sistem Information *Retrieval* mampu menjalankan proses pencarian dokumen sesuai *query* dengan melalui tahapan *preprocessing*, pembobotan TF-IDF, serta perhitungan *cosine similarity*. Evaluasi performa pada kedua threshold menunjukkan adanya perbedaan yang cukup jelas. Threshold 10 menghasilkan nilai *precision* yang tinggi, namun memiliki *recall* yang sangat rendah sehingga kurang mampu menemukan seluruh dokumen relevan. Sebaliknya, threshold 15 memberikan *recall* yang lebih besar sehingga berdampak pada peningkatan nilai *F1-measure* secara keseluruhan.

Nilai rata-rata *precision* 0.733, *recall* 0.037, dan *F1-measure* 0.070 pada threshold 15 menunjukkan bahwa konfigurasi ini merupakan yang paling optimal dalam konteks penelitian ini, karena memberikan keseimbangan terbaik antara ketepatan dan kelengkapan hasil pencarian. Dengan demikian, threshold 15 dapat direkomendasikan sebagai parameter yang paling sesuai untuk sistem pencarian dokumen pada dataset soal Sosiologi yang digunakan dalam penelitian ini.

## Daftar Pustaka

[1] M. Danuri, M. Informatika, J. Teknologi, and C. Semarang, “PERKEMBANGAN DAN TRANSFORMASI TEKNOLOGI DIGITAL,” 2019.

[2] B. Anjani, Y. Sugiarti, D. Lestari, R. Program, S. Pendidikan, and T. Agroindustri, “PENGEMBANGAN BANK SOAL DIGITAL INTERAKTIF PADA KOMPETENSI DASAR MENGANALISIS SIFAT BAHAN HASIL PERTANIAN,” 2019, [Online]. Available: http://ejournal.upi.edu/index.php/edufortech/indexEDUFORTECH4

[3] Ferry Sanjaya, “Pemanfaatan Sistem Temu Kembali Informasi dalam Pencarian Dokumen Menggunakan Metode Vector Space Model,” 2017.



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

[4] K. D. Putung, A. Lumenta, and A. Jacobus, "PENERAPAN SISTEM TEMU KEMBALI INFORMASI PADA KUMPULAN DOKUMEN SKRIPSI," *18 E-journal Teknik Informatika*, vol. 8, no. 1, 2016.

[5] D. Nugraha, "Penerapan Algoritma Cosine Similarity Pada Aplikasi Bank Soal 2021," 2021.

[6] Nanang Setiawan and Fatkhul Amin, "Sistem Temu Kembali Informasi Jurnal Ilmiah Unisbank Dengan Metode Cosine Similarity," 2024.

[7] A. H. Nasrullah, "Integrasi Tf-Idf Dan Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer UNISAN Gorontalo)," *INTEC Journal: Information Technology Education Journal*, vol. 3, no. 3, 2024, [Online]. Available: https://scholar.google.com/,

[8] R. Al Rasyid, D. Handayani, and U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, p. 2024, 2024, doi: 10.35870/jti.

[9] D. Septiani and I. Isabela, "SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia ANALISIS TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DALAM TEMU KEMBALI INFORMASI PADA DOKUMEN TEKS".

[10] Ernawati, "PERPUSTAKAAN DIGITAL DALAM TEMU KEMBALI INFORMASI DENGAN OPAC Ernawati," 2018. [Online]. Available: http://puslit2.petra.ac.id/ejournal/index.php/pus/article/download/17222

[11] D. W. T. PUTRA and J. J. PUTRA, "PERANCANGAN SISTEM INFORMASI PENCARIAN LOWONGAN PEKERJAAN," *JURNAL TEKNOIF*, vol. 6, no. 1, pp. 48–54, Apr. 2018, doi: 10.21063/jtif.2018.v6.1.48-54.

[12] S. Yusuf, M. A. Fauzi, and K. C. Brata, "Sistem Temu Kembali Informasi Pasal-Pasal KUHP (Kitab Undang-Undang Hukum Pidana) Berbasis Android Menggunakan Metode Synonym Recognition dan Cosine Similarity," 2018. [Online]. Available: http://j-ptiik.ub.ac.id



https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI