KLASIFIKASI JURNAL KESEHATAN DIGITAL MENGGUNAKAN TF-IDF DAN K-NEAREST NEIGHBOR

Neisya Nur Qoyimah, Otniel Silaban, Nur Khanifah, Laili Cahyani
Program Studi Pendidikan Informatika, Universitas Trunojoyo Madura
230631100066@student.trunojoyo.ac.id

ABSTRAK

Peningkatan jumlah publikasi ilmiah di bidang kesehatan digital dalam beberapa tahun terakhir menyebabkan proses penelusuran dan pengelompokan dokumen secara manual menjadi sulit dilakukan. Kondisi tersebut harus ada sistem yang mampu membantu proses pengolahan informasi secara cepat, efisien, dan akurat. Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi otomatis pada abstrak jurnal kesehatan digital menggunakan kombinasi metode TF-IDF dan algoritma *K-Nearest Neighbor* (KNN). Penelitian ini juga memberikan kontribusi dalam penerapan pelabelan otomatis berbasis *rule-based* untuk membantu proses klasifikasi abstrak jurnal kesehatan digital menggunakan kombinasi metode TF-IDF dan algoritma *K-Nearest Neighbor* (KNN), yang masih belum banyak diterapkan di penelitian lain. Dataset yang digunakan terdiri dari 410 abstrak jurnal yang diproses melalui tahapan preprocessing, meliputi case folding, pembersihan tanda baca, tokenisasi, penghapusan stopword, dan stemming. Representasi numerik dilakukan menggunakan TF-IDF yang menghasilkan 4.753 fitur unik. Dataset memiliki distribusi label yang cukup seimbang yaitu 225 data kesehatan digital (54,88%) dan 185 non-digital (45,12%), serta dibagi menggunakan metode hold-out dengan komposisi 80% data latih (328 data) dan 20% data uji (82 data). Proses klasifikasi menggunakan algoritma KNN dengan nilai K=3 dan metrik cosine similarity. Hasil evaluasi menunjukkan performa model dengan accuracy 0,7195, precision 0,6667, recall 0,7895, dan F1-score 0,7229. Analisis confusion matrix menunjukkan 29 True Negative, 30 True Positive, 15 False Positive, dan 8 False Negative. Penelitian ini menunjukkan bahwa kombinasi TF-IDF dan KNN efektif untuk klasifikasi awal dokumen ilmiah, meskipun kualitas label otomatis dan keterbatasan dataset masih menjadi faktor yang perlu diperhatikan dalam pengembangan selanjutnya.

Kata Kunci : *Jurnal Kesehatan digital, TF-IDF, KNN, Klasifikasi, Temu Kembali Informasi, Text Mining*

1. PENDAHULUAN

Kesehatan digital merupakan sebuah perubahan revolusioner dalam bidang layanan kesehatan yang mendorong batas-batas tradisional praktik medis. Perkembangan teknologi digital, perangkat lunak kesehatan, serta berbagai platform daring telah menghadirkan beragam inovasi yang membantu meningkatkan kualitas layanan, pemantauan kondisi pasien, serta manajemen penyakit secara lebih efektif.s Konsep ini kini menjadi aspek penting bagi seluruh lapisan masyarakat, tanpa memandang usia, kondisi sosial, maupun tingkat ekonomi [1]. Di era globalisasi dan transformasi digital, teknologi informasi telah menjadi bagian integral dalam penyelenggaraan layanan kesehatan [2]. Meningkatnya jumlah publikasi ilmiah dalam bidang teknologi dan kesehatan, proses pengelolaan dokumen menjadi semakin kompleks. Banyaknya artikel yang diterbitkan setiap tahun menyebabkan peneliti kesulitan dalam melakukan penelusuran, pengelompokan, dan identifikasi topik secara manual. Metode klasifikasi konvensional tidak lagi efektif karena membutuhkan waktu lama, rentan terhadap subjektivitas penilai, dan tidak mampu menangani volume data yang besar secara efisien [3].

Perkembangan layanan kesehatan digital didorong oleh berbagai faktor, salah satunya adalah peningkatan penggunaan internet serta perubahan perilaku masyarakat dalam mencari informasi kesehatan. Abdullah Syafei dalam Jurnal Ilmu Kesehatan Masyarakat menjelaskan bahwa masyarakat semakin memanfaatkan sumber digital untuk memperoleh informasi kesehatan. Secara global, lebih dari 5,16 miliar penduduk telah mengakses internet, dan di beberapa negara seperti Swiss, sekitar 60% responden menjadikan media digital sebagai sarana utama dalam memperoleh informasi terkait kesehatan. [4]. Hal ini menunjukkan bahwa transformasi digital semakin melekat dalam proses pengambilan keputusan kesehatan modern. Oleh karena itu, kesehatan digital tidak hanya menjadi inovasi teknologi, tetapi juga kebutuhan masyarakat.

Seiring dengan percepatan inovasi kesehatan digital, jumlah publikasi ilmiah terkait teknologi kesehatan juga meningkat. Dalam penelitian Abdullah Syafei proses penelusuran literatur dari 1798 artikel jurnal, didapatkan 7 artikel yang memenuhi kriteria setelah melalui proses seleksi ketat [5]. Data tersebut menunjukkan betapa besar volume publikasi di bidang kesehatan digital, sekaligus memperlihatkan kesulitan peneliti dalam menyaring dan mengidentifikasi artikel relevan secara manual. Kondisi ini memperkuat urgensi penggunaan metode klasifikasi otomatis untuk mempermudah pemetaan topik-topik ilmiah.

Besarnya jumlah data dan percepatan digitalisasi di sektor kesehatan membuat pengelolaan informasi menjadi semakin kompleks. Dalam penelitian oleh Fazila Septiani Santoso et al. (2025), transformasi digital dalam layanan kesehatan menghasilkan data yang besar dan beragam, sehingga diperlukan dukungan teknologi untuk meningkatkan efisiensi dan ketepatan pengelolaan informasi [6].

Penelitian sebelumnya juga telah membahas pemanfaatan teknologi digital dan kecerdasan buatan untuk meningkatkan kualitas layanan kesehatan, seperti penerapan AI untuk diagnosis citra dan telemedicine [7]. Selain itu, penggunaan metode TF-IDF dan algoritma KNN dengan cosine similarity pada klasifikasi abstrak jurnal kesehatan digital, khususnya yang menggunakan pelabelan otomatis, masih belum banyak dibahas pada penelitian sebelumnya. Dengan demikian, penelitian ini masih memiliki potensi untuk dikembangkan lebih lanjut dalam model klasifikasi otomatis yang mampu mengelompokkan dokumen ilmiah di bidang kesehatan digital secara efektif.

Beberapa penelitian sebelumnya lebih fokus pada klasifikasi dokumen umum atau penerapan AI di bidang kesehatan, namun belum spesifik dalam membahas klasifikasi abstrak jurnal kesehatan digital dengan pendekatan pelabelan otomatis berbasis rule-based. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model klasifikasi abstrak jurnal kesehatan digital menggunakan kombinasi TF-IDF dan KNN berbasis cosine similarity. Berdasarkan permasalahan tersebut, penelitian ini memiliki beberapa kontribusi utama, yaitu: (1) mengembangkan model klasifikasi abstrak jurnal kesehatan digital menggunakan kombinasi TF-IDF dan KNN berbasis cosine similarity, (2) menerapkan pelabelan otomatis berbasis *rule-based* sebagai pendekatan awal dalam proses anotasi data, serta (3) menganalisis performa model pada dataset abstrak jurnal dengan distribusi kelas yang cukup seimbang. Kontribusi ini diharapkan dapat memberikan pendekatan awal dalam pengembangan sistem klasifikasi dokumen ilmiah berbasis otomatis.

2. TINJAUAN PUSTAKA

2.1. Kesehatan Digital

Kesehatan digital merupakan pemanfaatan teknologi digital dalam layanan kesehatan, seperti eHealth, mHealth, telehealth atau telemedicine, rekam medis elektronik (EMR) [8]. Selain itu, kesehatan digital juga dapat mendukung pengelolaan data medis dalam jumlah besar yang dimanfaatkan untuk analisis dan pengambilan keputusan. Dalam penelitian ini, kesehatan digital digunakan sebagai dasar dalam proses pengelompokan abstrak jurnal menjadi kategori kesehatan digital dan non-digital.

2.2. Klasifikasi

Klasifikasi adalah suatu teknik pengelompokan dokumen ke dalam kategori kelas-kelas tertentu secara otomatis berdasarkan karakteristik atau ciri yang serupa. Proses ini bertujuan untuk memperoleh fungsi atau model yang dapat menjelaskan dan membedakan konsep atau kelas data, ketika model telah ditemukan maka model ini dapat digukanakan untuk melakukan prediksi kelas data baru [9].

2.3. Text Mining

Text Mining merupakan teknik dalam klasifikasi yang diterapkan untuk mengekstrak informasi yang relefan dari sekumpulan dokumen [10]. Teknik ini banyak diterapkan dalam proses klasifikasi dokumen untuk membantu mengatasi permasalahan pengelompokan data teks secara otomatis [11]. Dalam penelitian ini, text mining digunakan untuk mengolah abstrak jurnal sebelum dilakukan pembobotan TF-IDF dan proses klasifikasi menggunakan KNN.

2.4. Text Preprocessing

Text preprocessing merupakan sebuah teknik untuk mengolah teks dengan tujuan untuk membersihkan dan menyiapkan data teks mentah agar bisa diproses lebih lanjut oleh sistem. Terdapat 5 tahapan dalam text preprocessing, yakni [12]:

a. Case Folding
   Case Folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil.
b. Punctuation Removal
   Punctuation Removal merupakan proses penghapusan tanda baca yang tidak memiliki pengaruh penting terhadap proses analisis teks.
c. Tokenisasi
   Tokenisasi merupakan suatu proses pemecahan teks panjang menjadi potongan-potongan kata atau istilah yang lebih kecil. Tujuannya agar sistem dapat membaca dan memahami isi teks secara lebih terstruktur.
d. Stopword Removal
   Stopword removal adalah proses menghapus kata-kata yang terlalu umum dan tidak memberi pengaruh besar terhadap analisis, misalnya kata “dan”, “yang”, atau “dengan”.
e. Stemming
   Stemming adalah teknik mengubah kata ke bentuk dasarnya dengan menghilangkan kata imbuhan. seperti, kata “menggunakan” diubah menjadi “guna”.

2.5. TF (*Term Frequency*) – IDF (*Inverse Document Frequency*)

TF-IDF (*Term Frequency–Inverse Document Frequency*) merupakan teknik pembobotan kata yang digunakan dalam pengolahan teks untuk mengukur tingkat relevansi suatu kata pada sebuah dokumen dibandingkan dengan seluruh koleksi

dokumen. Metode ini mengombinasikan dua komponen utama, yaitu *Term Frequency (TF)* yang menghitung seberapa sering suatu kata muncul dalam satu dokumen [13]. Tujuan pembobotan kata ini agar data yang belum terstruktur menjadi lebih terstruktur.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

Keterangan:

a. *TF (t,d)* = nilai Term Frequency dari kata *t* pada dokumen *d*
b. $F\,t,d$ = jumlah kemunculan kata *t* dalam dokumen *d*
c. $\sum_k f\,k,d$ = total seluruh kata dalam dokumen *d*

Sedangkan *Inverse Document Frequency (IDF)* yang menilai seberapa jarang kata tersebut muncul di keseluruhan dokumen, sehingga kata yang sering muncul tetapi bersifat umum akan memiliki bobot rendah [13].

$$IDF(t) = log\frac{N}{df(t)}$$

Keterangan:

a. *IDF (t)* = nilai Inverse Document Frequency dari kata *t*
b. *N* = jumlah seluruh dokumen
c. $df(t)$ = jumlah dokumen yang mengandung kata t

Hasil dari TF-IDF berupa nilai numerik yang merepresentasikan dokumen dalam bentuk vektor, sehingga dapat digunakan untuk mengukur kemiripan antar dokumen secara matematis, khususnya dalam sistem rekomendasi jurnal berbasis teks [13].

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

Keterangan:

a. *TF-IDF (t,d)* = bobot kata *t* dalam dokumen *d*
b. *TF (t,d)* = nilai frekuensi kata
c. *IDF (t)* = nilai kebalikan frekuensi dokumen

2.6. Algoritma K-Nearest Neighbor (KNN)

Algoritma K-Nearest Neighbor (KNN) merupakan metode klasifikasi yang menentukan kategori suatu objek berdasarkan jumlah data pelatihan yang memiliki jarak paling dekat dengan objek tersebut. K-Nearest Neighbor didasarkan pada konsep 'learning by analogy'[14]. Dalam penelitian ini digunakan metrik cosine similarity untuk mengukur kedekatan antar dokumen, karena metrik ini lebih sesuai untuk data berdimensi tinggi hasil representasi TF-IDF dibandingkan Euclidean distance, sebab cosine similarity mengukur sudut antara dua vektor dokumen sehingga lebih robust terhadap perbedaan panjang dokumen.

Pemilihan nilai K=3 pada penelitian ini didasarkan pada beberapa pertimbangan. Nilai K=1 terlalu sensitif terhadap noise pada data, sedangkan nilai K yang lebih besar seperti K=5 atau K=7 cenderung mengikutsertakan tetangga yang kurang relevan secara konteks pada representasi TF-IDF yang berdimensi tinggi. Nilai K=3 dipilih sebagai titik keseimbangan antara stabilitas prediksi dan sensitivitas terhadap konteks dokumen.

2.7. Evaluation Metrics

Evaluasi menggunakan data uji dengan menerapkan beberapa metrik evaluasi, yaitu akurasi, presisi, recall, dan F1-score. Seluruh metrik tersebut dihitung berdasarkan komponen True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Selain itu, confusion matrix digunakan sebagai alat bantu untuk menganalisis kinerja model secara lebih rinci pada setiap kelas, sehingga dapat memberikan gambaran yang lebih jelas mengenai tingkat kesalahan dan keberhasilan klasifikasi yang dihasilkan oleh model.

a. Akurasi (*Accuracy*)
   Akurasi (*Accuracy*) merupakan ukuran seberapa tepat sebuah model dalam melakukan prediksi. Nilai ini dihitung dari jumlah prediksi yang benar dibandingkan dengan seluruh data yang diuji.
   $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
   Keterangan:
   a. TP = True Positive (prediksi positif yang benar)
   b. TN = True Negative (prediksi negatif yang benar)
   c. FP = False Positive (prediksi positif yang salah)
   d. FN = False Negative (prediksi negatif yang salah)
b. Presisi (*Precision*)
   Presisi (*Precision*) merupakan suatu teknik untuk menunjukkan seberapa banyak hasil prediksi positif yang benar-benar tepat. Metode ini membantu melihat apakah model tidak terlalu sering salah dalam memberikan prediksi positif.
   $$Precision = \frac{TP}{TP + FP}$$
   Keterangan:
   a. TP = True Positive
   b. FP = False Positive
c. Recall
   Recall untuk mengukur kemampuan model dalam menemukan data positif secara keseluruhan. Nilai recall tinggi berarti model berhasil mengenali hampir semua data yang seharusnya masuk kategori positif.
   $$Recall = \frac{TP}{TP + FN}$$
   Keterangan:
   a. TP = True Positive
   b. FN = False Negative
d. F1-score
   F1-score merupakan nilai gabungan antara presisi dan recall yang dirata-ratakan secara harmonik. Ukuran ini digunakan ketika diperlukan keseimbangan antara ketepatan

prediksi dan kemampuan model menemukan data positif.

$$F1\text{-}Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

Keterangan:

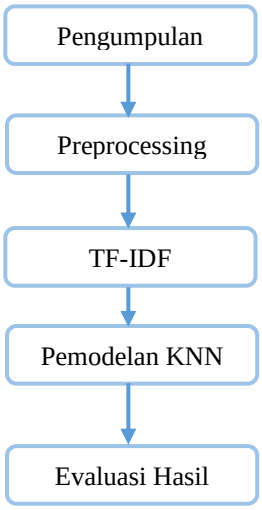
a. Precision = nilai presisi model
b. Recall = nilai recall model

e. Confusion matrix

Confusion matrix merupakan tabel yang memperlihatkan jumlah prediksi yang benar dan salah untuk setiap kelas. Melalui tabel ini, kita dapat melihat lebih jelas pola kesalahan model dan bagaimana performa prediksi pada tiap kategorinya.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan pengolahan teks *(text mining)* untuk melakukan klasifikasi pada abstrak jurnal kesehatan digital dengan memanfaatkan metode TF-IDF dan algoritma K-Nearest Neighbor (KNN). Proses penelitian dilakukan secara terstruktur mulai dari pengumpulan data hingga evaluasi model klasifikasi agar menghasilkan sistem yang akurat dan dapat mengelompokkan dokumen sesuai kategori. Terdapat beberapa tahapan yang dilakukan dalam penelitian ini sebagai berikut:



Gambar 1. Tahap Metode Penelitian

3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan mengambil 410 abstrak jurnal kesehatan digital dari berbagai sumber publikasi ilmiah di bidang kesehatan dan teknologi informasi. Seluruh abstrak kemudian dipilih sesuai kriteria penelitian dan disusun dalam bentuk dataset yang terstruktur. Data ini menjadi dasar untuk tahapan selanjutnya, yaitu preprocessing teks, perhitungan bobot TF-IDF, dan proses klasifikasi. Ringkasan data yang digunakan ditampilkan pada Tabel 1.

Tabel 1. Data abstrak jurnal kesehatan digital

No	Tahun	Abstrak
1.	2007	Mobilitas manusia pada era globalisasi...
2.	2012	Layanan kesehatan prima atau...
3.	2013	Klinik sebagai tempat yang...
4.	2014	Obat adalah bahan atau zat yang berasal...
5.	2015	Teknologi informasi saat ini menjadi...
6.	2015	Sistem Informasi rekam medis merupakan...
7.	2015	Penelitian ini merepresentasikan data...
8.	2015	Sistem pencarian merupakan salah satu...
9.	2015	Penelitian ini merepresentasikan data...
...	...	...
410	2025	Stres adalah respons tubuh terhadap...

3.2. Preprocessing

Pada tahap preprocessing, data abstrak dibersihkan terlebih dahulu agar teksnya rapi dan mudah diproses oleh sistem. Dan untuk memastikan setiap teks berada dalam format yang konsisten sehingga memudahkan proses ekstraksi fitur dan klasifikasi pada tahap berikutnya. Tahap-tahap preprocessing meliputi:

a. Case Folding

Case Folding berfungsi untuk mengubah seluruh teks menjadi huruf kecil atau lowercase.

Tabel 2. Proses Case Folding

Sebelum	Sesudah
Saat terjadi pandemi corona virus disease (COVID-19), layanan ...	saat terjadi pandemi corona virus disease (covid-19), layanan ...

b. Punctuation Removal

Punctuation Removal berfungsi untuk menghapus tanda baca seperti koma, titik, dan simbol lainnya.

Tabel 3. Punctuation Removal

Sebelum	Sesudah
saat terjadi pandemi corona virus disease (covid-19), layanan ...	saat terjadi pandemi corona virus disease covid19 layanan ...

c. Tokenisasi

Tokenisasi berfungsi untuk memecah teks menjadi token kata sehingga setiap kata dapat diproses secara individual.

Tabel 4. Tokenisasi

Sebelum	Sesudah
saat terjadi pandemi corona virus disease covid19 layanan ...	['saat', 'terjadi', 'pandemi', 'corona', 'virus', 'disease', 'covid19', 'layanan', ...]

d. Stopword Removal
Stopword Removal berfungsi untuk menghapus kata-kata umum yang tidak memiliki nilai informasi signifikan dalam proses analisis.

Tabel 5. Stopword Removal

Sebelum	Sesudah
['saat', 'terjadi', 'pandemi', 'corona', 'virus', 'disease', 'covid19', 'layanan', ...]	['pandemi', 'corona', 'virus', 'disease', 'covid19', 'layanan', ...]

e. Stemming
Stemming berfungsi untuk mengubah kata ke bentuk dasarnya menggunakan algoritma Sastrawi agar variasi kata dengan makna sama diseragamkan.

Tabel 6. Stemming

Sebelum	Sesudah
['pandemi', 'corona', 'virus', 'disease', 'covid19', 'layanan', ...]	['pandemi', 'corona', 'virus', 'disease', 'covid19', 'layan', ...]

3.3. Ekstraksi Fitur TF-IDF (*Term Frequency-Inverse Document Frequency*)

Pada tahap Ekstraksi Fitur, peneliti menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) untuk mengubah teks yang telah melalui tahap preprocessing ke dalam representasi numerik. TF-IDF bekerja dengan menghitung frekuensi kemunculan kata dalam sebuah dokumen (TF) dan mengalikan dengan nilai kebalikan frekuensi kata pada keseluruhan dokumen (IDF), sehingga kata yang sering muncul dalam satu dokumen namun jarang muncul di dokumen lain akan memiliki bobot lebih tinggi.

Proses perhitungan TF-IDF dilakukan menggunakan fungsi TfidfVectorizer dari library scikit-learn. Hasil ekstraksi menghasilkan sebanyak 4753 fitur unik, dengan matriks berdimensi 410 × 4753, yang berarti 410 dokumen dikonversi menjadi vektor dengan 4753 fitur pembobotan. Sebagian contoh fitur yang dihasilkan dari proses ekstraksi TF-IDF ditampilkan pada tabel berikut.

Tabel 7. Contoh Hasil Pelabelan Otomatis

No	Fitur
1.	Ab
2.	Abai
3.	Abc
4.	Abdi
5.	Abdomen
...	...

3.4. Pelabelan Otomatis Berbasis *Rule-Based*

Penelitian ini menggunakan metode *rule-based* untuk memberikan label otomatis pada setiap abstrak. Label diberikan berdasarkan keberadaan kata kunci seperti *"digital", "telemedicine", "rekam medis", "AI", "IoT"* "*Machine Learning*" dan sebagainya. Jika abstrak mengandung salah satu kata kunci tersebut, maka dokumen diberi label 1 (kesehatan digital), sedangkan dokumen lain diberi label 0 (kesehatan non-digital).

Penggunaan pelabelan otomatis berbasis *rule-based* membantu mempercepat proses pemberian label pada dataset tanpa dilakukan secara manual. Namun, penentuan label hanya bergantung pada kemunculan kata kunci tertentu dapat menyebabkan beberapa abstrak tidak sepenuhnya merepresentasikan konteks sebenarnya. Kondisi ini memungkinkan terjadinya ketidaksesuaian label pada beberapa data, yang pada akhirnya dapat mempengaruhi hasil klasifikasi yang diperoleh. Oleh karena itu, performa model pada penelitian ini juga dipengaruhi oleh kualitas label yang bergantung pada kelengkapan kata kunci dan proses pelabelan otomatis yang digunakan.

Tabel 8. Contoh 5 Fitur TF-IDF Pertama

No	Abstrak	Label
1.	Mobilitas manusia pada era…	1
2.	Layanan kesehatan prima…	1
3.	Klinik sebagai tempat layanan…	1
4.	Obat adalah bahan…	0
5.	Teknologi informasi saat ini…	1
...	...	...

3.5. Pemodelan Algoritma K-Nearest Neighbor (KNN)

Pada tahap pemodelan algoritma, sistem melakukan proses klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN), dengan mengukur tingkat kemiripan antar dokumen berdasarkan jarak (*distance metric*). Pada penelitian ini digunakan metrik *cosine similarity* dengan nilai K = 3 sebagai jumlah *neighbor* yang dipertimbangkan dalam proses pengambilan keputusan. Pemilihan nilai K = 3 ini bertujuan untuk menjaga dokumen agar tidak sensitif terhadap kemiripan, karena pada representasi TF-IDF, kedekatan antar dokumen sangat dipengaruhi oleh kemiripan kata kunci yang spesifik.

Dataset dalam penelitian ini dibagi menggunakan metode *hold-out* dengan proporsi 80% data latih (328 dokumen) dan 20% data uji (82 dokumen). Skema *hold-out* dipilih karena jumlah dataset yang relatif terbatas dengan distribusi kelas yang cukup seimbang, serta karena prosesnya lebih sederhana dan tidak membutuhkan komputasi intensif dibandingkan *k-fold cross-validation*, sehingga sesuai dengan tujuan pengujian performa awal model.

3.6. Evaluasi Model

Pada tahap evaluasi, dilakukan dengan membandingkan hasil prediksi model terhadap label aktual pada data uji pada dataset abstrak jurnal. Metrik evaluasi dalam penelitian ini meliputi accuracy, precision, recall, dan F1-score. Selain itu, confusion matrix digunakan untuk menganalisis

distribusi hasil prediksi model terhadap masing-masing kelas secara lebih rinci.

Proses evaluasi diterapkan pada seluruh data abstrak yang telah melalui tahapan preprocessing, ekstraksi fitur menggunakan Term Frequency–Inverse Document Frequency (TF-IDF), serta proses klasifikasi menggunakan algoritma KNN dengan nilai K sebesar 3 dan metrik cosine similarity. Hasil evaluasi ini menjadi dasar dalam menilai efektivitas model dalam mengklasifikasikan jurnal kesehatan digital.

Skema evaluasi *hold-out* digunakan karena jumlah dataset pada penelitian ini masih dalam kategori terbatas, dengan distribusi kelas yang cukup seimbang. Pembagian data menjadi data pelatihan (80%) dan data pengujian (20%) dianggap memadai untuk mengevaluasi performa awal model klasifikasi. Selain itu, metode *hold-out* dipilih karena proses penerapannya lebih sederhana dan tidak membutuhkan komputasi yang kompleks, sehingga sesuai dengan tujuan penelitian yang berfokus pada pengujian performa awal model klasifikasi.

4. HASIL DAN PEMBAHASAN

Hasil penelitian ini berupa sistem klasifikasi abstrak jurnal kesehatan digital yang bertujuan untuk mengelompokkan jurnal ke dalam kategori kesehatan digital dan non-digital. Sistem bekerja dengan memanfaatkan dataset abstrak jurnal ilmiah yang diproses melalui tahapan text preprocessing untuk menghasilkan data teks yang lebih terstruktur.

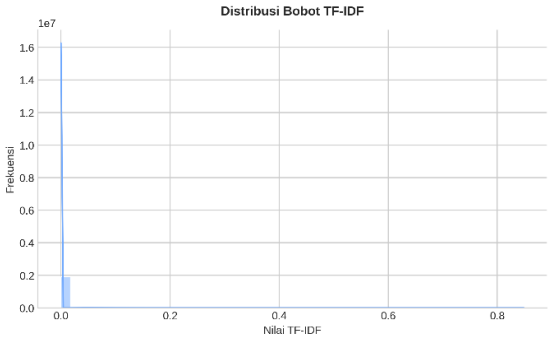
Tahapan text preprocessing berhasil membersihkan dan menormalkan seluruh 410 abstrak jurnal. Hasil preprocessing menunjukkan bentuk teks yang lebih terstruktur dan mengurangi kata-kata yang tidak penting dalam proses klasifikasi. Berikut hasil Perbandingan teks sebelum dan sesudah preprocessing ditampilkan pada Tabel 9.

Tabel 9. Hasil Tahap Preprocessing

Sebelum	Sesudah
Mobilitas manusia pada era globalisasi saat ini ...	['mobilitas', 'manusia', 'era', 'globalisasi', ...]
Layanan kesehatan prima atau excellent services ...	['layan', 'sehat', 'prima', 'excellent', 'services', ...]
Klinik sebagai tempat yang menyediakan layanan kesehatan ...	['klinik', 'sedia', 'layan', 'sehat', ...]
...	...
World Health Organization (WHO, 2010) melaporkan limbah	['world', 'health', 'organization', 'who2010', 'lapor', 'limbah', ...]
Saat terjadi pandemi corona virus disease (COVID-19), layanan ...	['pandemi', 'corona', 'virus', 'disease', 'covid19', 'layan', ...]
Stres adalah respons tubuh terhadap tekanan atau tantangan ...	['stres', 'respons', 'tubuh', 'tekan', 'tantang', …]

Setelah tahap preprocessing, setiap abstrak jurnal direpresentasikan dalam bentuk bobot kata menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Proses ini bertujuan untuk memberikan bobot pada setiap kata berdasarkan tingkat kemunculannya dalam dokumen dan keseluruhan korpus.

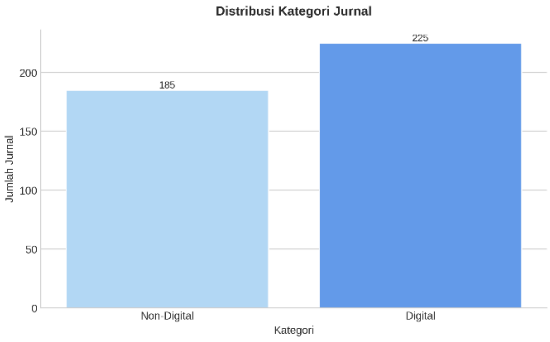
Hasil ekstraksi fitur TF-IDF pada penelitian ini menghasilkan representasi numerik seluruh abstrak dalam matriks berukuran 410 x 4.753, yang berarti setiap dokumen direpresentasikan oleh 4.753 fitur pembobotan kata unik. Berikut ini hasil distribusi nilai bobot TF-IDF ditampilkan pada Gambar 1.



Gambar 2. Distribusi Bobot TF-IDF

Distribusi bobot TF-IDF pada Gambar 1 menunjukkan bahwa sebagian besar nilai mendekati nol, yang merupakan karakteristik umum dari representasi TF-IDF pada data teks. Kondisi ini menggambarkan tingginya sparsity pada matriks fitur, di mana setiap dokumen hanya mengandung sebagian kecil dari total 4.753 kata unik yang ada dalam seluruh korpus. Kata-kata yang memiliki bobot TF-IDF tinggi merupakan kata-kata yang spesifik pada dokumen tertentu dan jarang muncul di dokumen lain, sehingga menjadi fitur diskriminatif yang paling berpengaruh dalam proses klasifikasi.

Hasil pelabelan otomatis terhadap abstrak jurnal berbasis rule-based menghasilkan distribusi 225 abstrak berlabel 1 (kesehatan digital, 54,88%) dan 185 abstrak berlabel 0 (non-digital, 45,12%). Distribusi ini ditampilkan pada Gambar 2.



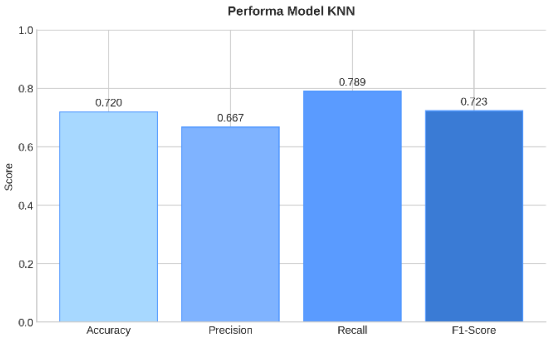
Gambar 3. Distribusi Kategori Jurnal

Distribusi kelas yang relatif seimbang (54,88% vs 45,12%) merupakan kondisi yang menguntungkan dalam proses pelatihan model, karena model tidak terlalu terdominasi oleh salah satu kelas sehingga metrik seperti accuracy dapat mencerminkan performa aktual secara lebih proporsional.

Hasil klasifikasi yang dilakukan menggunakan algoritma K-Nearest Neighbor (KNN) dengan konfigurasi cosine similarity dan nilai K sebesar 3. Berikut ini hasil metrik evaluasi sebagaimana ditampilkan pada Tabel 10 dan Gambar 3.

Tabel 10. Hasil Evaluasi Model *K-Nearest Neighbor* (KNN)

Metrik	Nilai
Accuracy	0,7195
Precision	0,6667
Recall	0,7895
F1-Score	0,7229

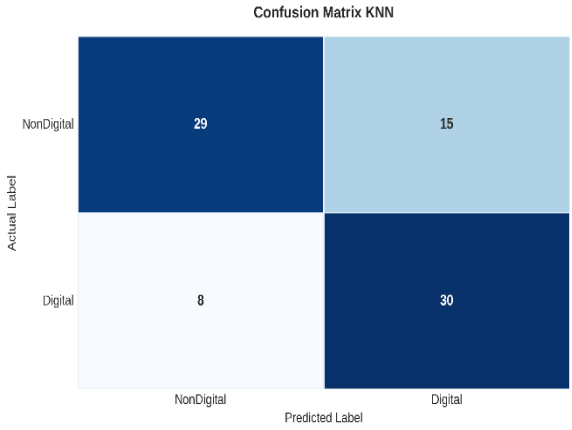


Gambar 4. Performa Model *K-Nearest Neighbor* (KNN)

Berdasarkan hasil evaluasi, nilai accuracy sebesar 0,7195 menunjukkan bahwa model mampu memberikan prediksi yang benar pada 71,95% dari seluruh data uji. Nilai precision sebesar 0,6667 mengindikasikan bahwa sekitar 66,67% prediksi positif (kesehatan digital) yang dihasilkan model merupakan prediksi yang benar. Nilai recall sebesar 0,7895 menunjukkan bahwa model berhasil mengidentifikasi 78,95% dari seluruh dokumen yang sebenarnya termasuk dalam kategori kesehatan digital.

Perbedaan antara nilai recall (0,7895) dan precision (0,6667) mengindikasikan bahwa model cenderung lebih sensitif dalam mendeteksi kelas positif, namun menghasilkan sejumlah false positive. Kondisi ini dapat dipahami mengingat data latih memiliki proporsi kelas positif yang sedikit lebih dominan (57,01% label digital pada data latih), sehingga model memiliki tendensi untuk memprediksi kelas positif lebih sering. Nilai F1-score sebesar 0,7229 mencerminkan keseimbangan yang cukup baik antara precision dan recall secara keseluruhan.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model KNN mampu melakukan klasifikasi abstrak jurnal kesehatan digital dengan performa yang cukup baik. Nilai accuracy 0,7195 mencerminkan bahwa lebih dari 70% prediksi model sudah benar, sementara nilai F1-score 0,7229 menunjukkan keseimbangan yang memadai antara precision dan recall. Meskipun demikian, masih terdapat ruang peningkatan terutama dalam mengurangi kesalahan klasifikasi pada masing-masing kelas.



Gambar 5. Confusion Matrix *K-Nearest Neighbor* (KNN)

Tabel 11. Detail Distribusi Prediksi Model

Kelas Aktual	Prediksi Digital	Prediksi Non-Digital
Digital	30 (True Positive)	8 (False Negative)
Non-Digital	15 (False Positive)	29 (True Negative)

Hasil ini diperkuat oleh confusion matrix yang menunjukkan bahwa model menghasilkan nilai True Positive (TP) sebesar 30 dan True Negative (TN) sebesar 29, yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Selain itu, terdapat False Positive (FP) sebanyak 15 dan False Negative (FN) sebanyak 8, nilai FP yang lebih besar dibandingkan nilai FN, kondisi ini menunjukkan bahwa masih terdapat beberapa kesalahan klasifikasi pada kedua kelas. Kesalahan klasifikasi yang terjadi dapat disebabkan oleh kemiripan kata antar dokumen, serta penggunaan pelabelan otomatis berbasis kata kunci yang masih memiliki keterbatasan dalam merepresentasikan konteks secara menyeluruh. Meskipun demikian, kesalahan tersebut masih dalam batas yang wajar dan tidak mengurangi kemampuan model secara signifikan dalam melakukan klasifikasi.

Meskipun demikian, hasil akurasi model pada penelitian ini (71,95%) berada pada rentang yang sebanding dengan penelitian sejenis yang menggunakan kombinasi TF-IDF dan KNN untuk

klasifikasi dokumen teks. Wijaya et al. (2025) melaporkan penerapan TF-IDF dan KNN pada sistem rekomendasi jurnal otomatis dengan hasil yang cukup baik [13]. Perbedaan performa antar penelitian umumnya dipengaruhi oleh karakteristik domain dataset, ukuran dataset, kualitas label, dan nilai K yang digunakan. Pada penelitian ini, penggunaan pelabelan otomatis berbasis rule-based sebagai pengganti anotasi manual merupakan faktor yang berpotensi membatasi performa model dibandingkan penelitian yang menggunakan label manual, karena kualitas label secara langsung mempengaruhi pola yang dipelajari model KNN.

5. KESIMPULAN DAN SARAN

Penelitian ini mengembangkan model klasifikasi abstrak jurnal kesehatan digital menggunakan kombinasi TF-IDF dan *K-Nearest Neighbor* (KNN) berbasis cosine similarity, dengan pelabelan otomatis berbasis *rule-based* sebagai pendekatan anotasi data. Dari data 410 abstrak jurnal yang digunakan, proses ekstraksi fitur TF-IDF menghasilkan 4.753 fitur unik, sedangkan pelabelan otomatis menghasilkan distribusi 225 data kesehatan digital dan 185 non-digital. Model *K-Nearest Neighbor* (KNN) dengan K=3 dan skema evaluasi hold-out (80% data latih, 20% data uji) menghasilkan accuracy 0,7195, precision 0,6667, recall 0,7895, F1-score 0,7229.

Analisis confusion matrix menunjukkan bahwa model menghasilkan 30 *True Positive*, 29 *True Negative*, 15 *False Positive*, dan 8 *False Negative.* Nilai FP yang lebih tinggi dari FN mengindikasikan kecenderungan model untuk memprediksi kelas positif lebih sering, yang konsisten dengan dominasi kelas digital pada data latih dan keterbatasan pelabelan otomatis berbasis kata kunci.

Terdapat beberapa keterbatasan yang perlu diakui dalam penelitian ini. Pertama, pelabelan otomatis berbasis rule-based bergantung sepenuhnya pada kelengkapan daftar kata kunci yang ditentukan, sehingga berpotensi menghasilkan label yang tidak sepenuhnya akurat. Kedua, dataset yang digunakan masih relatif terbatas (410 dokumen) sehingga representativitas model terhadap seluruh domain publikasi kesehatan digital belum dapat dijamin. Ketiga, penelitian ini hanya menguji satu algoritma klasifikasi sehingga perbandingan performa lintas algoritma belum dapat dilakukan.

Saran untuk pengembangan selanjutnya, penelitian ini masih memiliki ruang untuk diperbaiki. Dataset dapat diperluas dengan menambah jumlah abstrak agar distribusi data lebih representatif. Selain itu, pelabelan otomatis dapat diganti atau dilengkapi dengan anotasi manual untuk meningkatkan kualitas label. Variasi nilai K pada algoritma KNN juga perlu diuji lebih lanjut untuk memperoleh konfigurasi yang lebih optimal. Penggunaan metode representasi teks berbasis word embedding atau transformer dapat dipertimbangkan untuk memperkaya fitur teks. Selain itu, perbandingan performa model KNN dengan algoritma klasifikasi lain seperti Naive Bayes, SVM, atau Random Forest perlu dilakukan sebagai pembanding performa model.

DAFTAR PUSTAKA

[1] E. P. P. & K. Andryanto A., *Teknologi Kesehatan Digital*. 2025. [Online]. Available: https://penerbiteureka.com/2025/02/25/teknologi-kesehatan-digital/

[2] F. A. & Y. Y. Ridho Akbar Fadilah, "Perancangan Sistem Informasi Konsolidasi Integritas Rekam Medis Rawat Jalan BerbasisWeb Guna Menunjang Efektivitas Pengklaiman BPJS di Rumah Sakit Hermina Arcamanik Bandung," *J. Indones. Manaj. Inform. dan Komun.*, vol. 5, no. 2, pp. 1921–1937, 2024, [Online]. Available: https://journal.stmiki.ac.id/index.php/jimik/

[3] L. B. & R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, pp. 1–28, 2015, [Online]. Available: https://doi.org/10.1002/asi.23329

[4] C. M. Annur, "Jumlah Pengguna Internet Global Tembus 5,16 Miliar Orang pada Januari 2023.," Databoks. [Online]. Available: https://databoks.katadata.co.id/teknologi-telekomunikasi/statistik/3e28faffa14157a/jumlah-pengguna-internet-global-tembus-516-miliar-orang-pada-januari-2023

[5] A. Syafei, "Literasi Kesehatan Digital, Faktor yang Mempengaruhi, dan Hubungannya dengan Perilaku Kesehatan: Scoping Review," *J. Ilmu Kesehat. Masy.*, vol. 12, no. 50, pp. 545–553, 2023, [Online]. Available: https://doi.org/10.33221/jikm.v12i06.3232

[6] D. A. & S. H. Fazila Septiani Santoso, Putri Aulia Ramadhani and Purba, "Transformasi Digital Dalam Sektor Kesehatan Kajian Literatur Untuk Mendukung Inovasi dan Efisiensi Layanan Kesehatan," *Cindoku J. Keperawatan dan Ilmu Kesehat.*, vol. 2, no. 1, pp. 1–12, 2025, [Online]. Available: https://doi.org/10.61492/cindoku.v2i1.240

[7] M. F. Abdillah, "Revolusi Dgital Kesehatan: Meningkatkan Layanan dengan Kecerdasan Buatan," *Syntax Lit. J. Ilm. Indones.*, vol. 9, no. 10, pp. 5911–5921, 2024, [Online]. Available: https://doi.org/10.36418/syntax-literate.v9i10.50093

[8] H. Ni Ketut Yuliastuti, Juliatin, Siti Nur Janna, Eka Novia Syah Putri and A. & S. Risky, "Peran Digital Health dalam Peningkatan Akses dan Mutu Kesehatan di Indonesia: Tinjauan Literatur," *Preprotif J. Kesehat. Masy.*, vol. 10, no. 1, pp. 276–282, 2026,

[Online]. Available: https://doi.org/10.31004/prepotif.v10i1.54647

[9] R. Nanda, E. Haerani, S. K. Gusti, and S. Ramadhani, "Klasifikasi Berita Menggunakan Metode Support Vector Machine," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 2, pp. 269–278, 2022, doi: 10.32672/jnkti.v5i2.4193.

[10] I. & A. M. A. Sukriadi, "Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Yang Diusulkan Mahasiswa Menggunakan Metode Naïve Bayes," *J. Ilm. Sist. Inf. dan Tek. Inform.*, vol. 6, no. 2, pp. 184–196, 2023, [Online]. Available: https://doi.org/10.57093/jisti.v6i2.174

[11] E. Sabna, "Penerapan Text Mining untuk Pengelompokan Penelitian Dosen," *J. Ilmu Komput.*, vol. 9, no. 2, pp. 161–164, 2020, [Online]. Available: https://doi.org/10.33060/JIK/2020/Vol9.Iss2.183

[12] L. B. & N. A. A. Master Edison Siregar, "Pengembangan Sistem Pencarian Resep Makanan dengan Implementasi Text Preprocessing dan BM25," *J. Nas. Komutasi dan Teknol. Inf.*, vol. 8, no. 3, pp. 1458–1465, 2025, [Online]. Available: https://doi.org/10.32672/jnkti.v8i3.9137

[13] H. I. & A. R. Jonathan Wijaya, Fernando Sugianto Putra, "Implementasi TF_IDF dan KNN pada Rekomendasi Jurnal Otomatis," *J. Informatics Comput. Eng. Res.*, vol. 2, no. 1, pp. 18–24, 2025, [Online]. Available: https://doi.org/10.31963/jicer.v2i1.5565

[14] U. Malikussaleh, "SENASTIKA Universitas Malikussaleh," pp. 1–14, 2007.