Jurnal Informatika
E-ISSN: 2528-0090
P-ISSN: 2252-9160

Vol. 15, No. 1, Juni 2026, Page. 153~166
https://ejournal.unidayan.ac.id/index.php/JIU
DOI: https://doi.org/10.55340/jiu.v15i1.2719

Model Rekomendasi Konten Edukasi Diabetes pada Instagram dengan Integrasi TF-IDF dan *Cosine Similarity* berbasis *Natural Language Processing*

A Model for Recommending Diabetes Educational Content on Instagram Using TF-IDF and Cosine Similarity Based on Natural Language Processing

Anggita Risqi Nur Clarita*[1], Muhamad Fatchan[2], Karina Imelda[3]
[123] Program Studi Teknik Informatika, Universitas Pelita Bangsa, Indonesia
Email: [1]anggitarisqi312210450@mhs.pelitabangsa.ac.id

Article Info:	Received	Revised	Accepted	Published:
	05 Juni 2026	06 Juni 2026	17 Juni 2026	19 Juni 2026

Abstrak

Meningkatnya prevalensi Diabetes Melitus menuntut penguatan literasi kesehatan digital mandiri. Namun, mayoritas edukasi di Instagram bersifat generalis dan belum tersegmentasi sesuai karakteristik medis pasien. Penelitian ini bertujuan mengembangkan model rekomendasi konten edukasi diabetes personal menggunakan pendekatan Content-Based Filtering. Berbeda dari model pencocokan kata konvensional terdahulu yang gagal menangani tingginya tumpang tindih istilah medis pada teks non-formal, kebaruan penelitian ini terletak pada integrasi knowledge-base terminologi medis hierarkis (Tipe 1, Tipe 2, Umum) yang diselaraskan bersama praktisi kesehatan untuk memandu akurasi pembobotan. Metode yang digunakan meliputi Term Frequency-Inverse Document Frequency (TF-IDF) untuk representasi fitur dan Cosine Similarity untuk mengukur kemiripan leksikal antar-vektor teks. Hasil evaluasi terhadap 75 data uji menunjukkan capaian akurasi klasifikasi back-end sebesar 84% (Interval Kepercayaan 95%: [75,70%, 92,30%]), dengan kualitas urutan pemeringkatan rekomendasi terarah pada nilai rata-rata Precision at 3 (P@3) sebesar 88%. Analisis data empiris mengonfirmasi kesenjangan di lapangan di mana 72,45% konten didominasi materi Umum. Kesimpulannya, model ini menunjukkan potensi performa baik dalam meminimalkan bias prediksi pada dataset terkait, meski masih memiliki keterbatasan overlap kata kunci pada kategori Tipe 2 (Recall 0,76). Pengembangan ke depan memerlukan arsitektur multimodal untuk memproses informasi dari media visual.

***Kata Kunci**: Content-Based Filtering, Cosine Similarity, Diabetes, Instagram, TF-IDF.*

Jurnal Informatika Vol. 15, No. 1, Juni 2026, Page. 153~166
E-ISSN: 2528-0090 https://ejournal.unidayan.ac.id/index.php/JIU
P-ISSN: 2252-9160 DOI: https://doi.org/10.55340/jiu.v15i1.2719

Abstract

The rising prevalence of diabetes mellitus calls for the strengthening of self-directed digital health literacy. However, most educational content on Instagram is general in nature and has not been tailored to patients' specific medical characteristics. This study aims to develop a personalized diabetes educational content recommendation model using a content-based filtering approach. Unlike previous conventional keyword-matching models that failed to handle the high overlap of medical terms in informal texts, the novelty of this study lies in the integration of a hierarchical medical terminology knowledge base (Type 1, Type 2, General) aligned with healthcare practitioners to guide weighting accuracy. The methods used include Term Frequency-Inverse Document Frequency (TF-IDF) for feature representation and Cosine Similarity to measure lexical similarity between text vectors. Evaluation results on 75 test data points show a back-end classification accuracy of 84% (95% Confidence Interval: [75.70%, 92.30%]), with the quality of the ranked order of directed recommendations at an average Precision at 3 (P@3) of 88%. Empirical data analysis confirmed a gap in the field where 72.45% of the content is dominated by General material. In conclusion, this model demonstrates the potential for good performance in minimizing prediction bias on the relevant dataset, although it still has limitations regarding keyword overlap in Type 2 categories (Recall 0.76). Future development requires a multimodal architecture to process information from visual media.

Keywords: *Content-Based Filtering, Cosine Similarity, Diabetes, Instagram, TF-IDF.*

1. PENDAHULUAN

Penyakit *Diabetes Melitus* (DM) merupakan tantangan kesehatan global kronis yang menuntut kepatuhan manajemen mandiri (*self-care*) ketat dari penderita untuk mencegah komplikasi serius [1], [2]. Di era digital, media sosial menjadi poros utama diseminasi edukasi kesehatan masyarakat [1]. Namun, pesatnya arus informasi ini memicu fenomena infodemi (*infodemic*), yaitu banjir misinformasi medis yang tidak tervalidasi di ruang siber [3]. Kondisi ini sangat membahayakan penderita diabetes yang membutuhkan akurasi panduan terapi secara mutlak [4]. Oleh karena itu, diperlukan instrumen penapis informasi otomatis yang andal untuk menyaring konten edukasi pada *platform* digital.

Di antara berbagai *platform* yang tersedia, Instagram menonjol sebagai media berbasis visual dan teks pendek dengan tingkat keterikatan (*engagement rate*) tertinggi [3], [5], [6]. Secara teknis, *platform* ini memadukan infografis visual, video pendek (*Reels*), dan deskripsi tekstual (*caption*) secara simultan untuk menurunkan beban kognitif masyarakat dalam mencerna materi kesehatan yang kompleks [3]. Namun, *caption* Instagram menyajikan struktur data tekstual non-formal yang sarat akan derau (*noise*) berupa bahasa gaul (*slang*), singkatan, emotikon, hingga redundansi tagar (*hashtag*). Karakteristik data yang tidak terstruktur ini memerlukan pendekatan komputasi khusus untuk mengekstraksi informasi medis murni, terlebih karena mayoritas konten edukasi diabetes saat ini masih disajikan dalam format generalis [5]. Konten-konten tersebut belum tersegmentasi secara spesifik berdasarkan kebutuhan klinis riil pengguna, yang secara mendasar terbagi menjadi karakteristik Diabetes Tipe 1 (autoimun/ketergantungan insulin) dan Diabetes Tipe 2 (resistensi insulin/manajemen gaya hidup) [1], [2], [7].

Untuk mengolah data teks tidak terstruktur berskala besar pada media sosial tersebut, pemanfaatan kecerdasan buatan melalui *Natural Language Processing* (NLP) menjadi mutlak diperlukan. NLP memungkinkan sistem komputer memahami dan menginterpretasikan bahasa alami manusia secara kontekstual, yang telah terbukti andal dalam menyelesaikan masalah klasifikasi dan ekstraksi fitur teks [8], [9]. Kemampuan ekstraksi NLP ini membuka peluang integrasi ke dalam sistem rekomendasi berbasis konten (*Content-Based Filtering*). Secara fundamental, metode ini bekerja dengan menganalisis kemiripan fitur tekstual dokumen di dalam basis data terhadap profil preferensi klinis pengguna [10], [11], [12]. *Content-Based Filtering* merupakan sebuah metode penyaringan informasi yang bekerja dengan cara menganalisis kemiripan fitur atau karakteristik antara konten yang ada di dalam basis data dengan profil preferensi historis pengguna [13]

Di dalam arsitektur ini, metode *Term Frequency-Inverse Document Frequency* (TF-IDF) diterapkan untuk menaikkan bobot terminologi klinis yang langka sekaligus mengeliminasi kata umum non-informatif [14], [15]. Setelah fitur teks ditransformasikan menjadi vektor numerik, tingkat kedekatan semantik antar-dokumen diukur menggunakan *Cosine Similarity* karena efisiensinya dalam menghitung kecocokan arah vektor di ruang berdimensi tinggi tanpa terpengaruh variasi panjang teks [16], [17].

Beberapa penelitian terdahulu telah berupaya mengoptimalkan kinerja kombinasi algoritma tersebut, namun masih menyisakan batas metodologis yang belum terpecahkan. Penelitian pada [14] berhasil membuktikan keandalan model dalam menghasilkan akurasi tinggi pada klasifikasi domain buku berdasarkan suasana hati pembaca, tetapi efektivitas tersebut belum diuji maupun diadaptasi pada domain kesehatan yang memerlukan presisi informasi. Sementara itu, penelitian pada [18] menguji sistem rekomendasi pada ranah kursus berbasis kecerdasan buatan, namun eksperimen mereka sepenuhnya bersandar pada data formal yang terstruktur sehingga tidak dirancang untuk menangani tingginya gangguan kebahasaan (*noise*) serta singkatan non-formal yang mendominasi media sosial. Di sisi lain, meskipun penelitian pada [19] mengeksplorasi optimasi *Content-Based Filtering* menggunakan *Cosine Similarity* pada program pelatihan dan sukses mencatatkan tingkat kepresisian rata-rata sebesar 88%, model tersebut tidak dilengkapi dengan komponen *knowledge-base* terminologi medis yang rigid sehingga tidak mampu memisahkan item dokumen yang memiliki tingkat tumpang tindih (*overlap*) fitur kata kunci yang sangat tinggi. Hingga saat ini, belum ditemukan penelitian yang secara khusus mengintegrasikan metode TF-IDF, *Cosine Similarity*, dan *knowledge-base* terminologi medis hierarkis untuk kebutuhan personalisasi konten edukasi diabetes pada *platform* Instagram.

Ketidakmampuan algoritma pencocokan kata konvensional dalam menghadapi tumpang tindih (*overlap*) istilah pada teks non-formal media sosial dapat berdampak fatal secara klinis. Sebagai contoh, kata "insulin" atau "diet" sering kali muncul secara acak baik pada konten Diabetes Tipe 1 maupun Tipe 2. Tanpa adanya pemandu berupa batasan aturan (*rules*) yang *rigid*, sistem berisiko salah merekomendasikan konten terapi Tipe 1 kepada penderita Tipe 2 yang sebenarnya membutuhkan manajemen gaya hidup. Oleh karena itu, diperlukan integrasi struktur *knowledge-base* terminologi medis hierarkis guna memandu akurasi pembobotan fitur teks dan mengatasi ambiguitas leksikal tersebut pada *platform* Instagram.

Berdasarkan kesenjangan penelitian (*research gap*) tersebut, rumusan masalah dalam penelitian ini difokuskan pada bagaimana merancang sebuah model komputasi yang mampu mengatasi tingginya ambiguitas leksikal dan interferensi derau (*noise*) tekstual pada media sosial

Jurnal Informatika Vol. 15, No. 1, Juni 2026, Page. 153~166
E-ISSN: 2528-0090 https://ejournal.unidayan.ac.id/index.php/JIU
P-ISSN: 2252-9160 DOI: https://doi.org/10.55340/jiu.v15i1.2719

demi menghasilkan personalisasi rekomendasi yang akurat secara klinis. Kebaruan (*novelty*) ilmiah yang ditawarkan hadir sebagai solusi langsung atas kelemahan model pencocokan kata konvensional terdahulu, yaitu melalui pembangunan komponen *knowledge-base* (basis pengetahuan) terminologi medis hierarkis (Tipe 1, Tipe 2, dan Edukasi Umum) yang disusun berdasarkan literatur kesehatan serta diselaraskan bersama praktisi kesehatan untuk memandu akurasi pembobotan fitur teks. Selain itu, kebaruan diperkuat melalui penyematan algoritma ringkasan ekstraktif *LexRank* untuk memadatkan informasi medis bagi efisiensi kognitif pengguna awam. Artikel ini menyajikan formulasi matematis, arsitektur sistem, hasil implementasi aplikasi, grafik performa, serta analisis *trade-off* antara keamanan konten medis dan efisiensi komputasi model.

2. METODE PENELITIAN

Penelitian ini merupakan jenis penelitian terapan (*applied research*) yang dikombinasikan dengan desain eksperimental komputasional. Fokus utama penelitian ini adalah mengimplementasikan arsitektur *healthcare recommender system* untuk personalisasi literasi medis penderita diabetes secara otomatis. Pendekatan komputasi yang diterapkan berbasis pada *Content-Based Filtering* (CBF) dengan mengintegrasikan teknik *Natural Language Processing* (NLP) untuk ekstraksi fitur semantik *caption* Instagram. Pustaka utama yang mendasari pengembangan model ini mengacu pada standarisasi pembobotan teks berbasis frekuensi inversi dokumen dan pengukuran jarak kosinus sudut vektor berdimensi tinggi. Secara sistematis, tahapan eksperimen komputasional ini dimulai dari pengumpulan data via *web scraping*, pembersihan derau bahasa (*text preprocessing*), pembobotan kata TF-IDF, penyelarasan basis pengetahuan *query*, hingga penapisan teks ringkasan ekstraktif menggunakan algoritma *Lexrank*.

2.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini bersifat statis, yang dikumpulkan melalui teknik *web scraping* menggunakan pustaka *Selenium*. Pengambilan data dilakukan pada Oktober 2025 terhadap 13 akun Instagram edukasi kesehatan, seperti @halodoc, @klikdiabetes, @mganikcare, @diabetes.id, dan akun relevan lainnya. Proses ini menghasilkan sebanyak 732 postingan mentah yang kemudian disimpan dalam format CSV. Skenario akuisisi data non-formal dari media sosial ini mengacu pada metodologi pengumpulan korpus teks tidak terstruktur untuk kebutuhan *healthcare recommendation system* berskala besar [20]. Struktur *dataset* mentah tersebut mencakup atribut utama hasil *scraping* yang disajikan secara ringkas pada Tabel 1 berikut:

Tabel 1. Struktur *Dataset*

Nama Kolom	Deskripsi	Contoh Nilai
account	Nama akun Instagram sumber postingan.	perkeni_ig
url	Alamat URL penuh menuju postingan di Instagram.	https://www.instagram.com/p/DNxh1QkUtos/
caption	Teks lengkap unggahan (keterangan/deskripsi postingan).	#baca*caption* Rekomendasi Penyuntikan Insulin yang Benar! Sudah rutin …
hashtags	Daftar *hashtag* yang digunakan dalam postingan.	#baca*caption*, #suntikinsulin, #guladarah, #pasiendiabetes ...
likes	Jumlah *likes* atau *reactions* pada postingan.	(*hidden or reels*) atau 3.
tanggal	Tanggal dan waktu postingan diunggah.	2025-08-25T10:01:15.000Z

Dataset dalam format CSV ini selanjutnya digunakan sebagai input utama pada tahap *preprocessing* teks, pembobotan TF-IDF, perhitungan *Cosine similarity*, serta evaluasi performa model rekomendasi konten edukasi diabetes.

2.2 Pemrosesan Awal Teks *(Text Preprocessing)*

Natural Language Processing (NLP) merupakan cabang kecerdasan buatan yang fokus pada interaksi antara komputer dan bahasa alami manusia, dengan tujuan agar sistem dapat mengenali, memahami, serta memproses teks secara kontekstual [21]. Langkah ini sangat krusial dalam korpus media sosial karena karakteristik teksnya yang non-formal dan dipenuhi bias linguistik [22]. Mengacu pada standar arsitektur *pipeline Natural Language Processing* (NLP) untuk korpus kesehatan digital, tahapan pemrosesan awal ini mencakup lima langkah berurutan [20]:

1. *Case Folding*: Menyeragamkan seluruh karakter huruf menjadi *lowercase* agar kata yang sama memiliki representasi tunggal tanpa terpengaruh kapitalisasi.
2. *Cleansing*: Membersihkan teks dari elemen non-informatif khas media sosial secara berurutan, meliputi URL, *mention*, *hashtag*, emoji, angka, tanda baca, dan *whitespace* berlebih.
3. *Stopword Removal*: Menghilangkan kata-kata umum yang tidak memiliki signifikansi semantik (seperti kata sambung dan kata depan) menggunakan kamus komprehensif dari NLTK, Sastrawi, serta daftar *custom* istilah media sosial.
4. *Filtering* Edukatif: Proses penyaringan bertahap untuk mengekstraksi konten edukasi diabetes murni agar sistem dapat direplikasi. Penyaringan dilakukan dua tahap; tahap pertama (*Inclusion Filter*) mendeteksi teks berdasarkan 15 kata kunci medis utama seperti '*diabetes'*, '*gula'*, '*edukasi'*, '*pola makan*', '*insulin*"). Tahap kedua (*Exclusion Filter*) mengeliminasi konten promosi komersial atau acara dengan mendeteksi daftar kata pengecualian (*exclude words*) seperti *'seminar', 'webinar', 'promo', 'harga',* dan *'quiz'*. Konten yang lolos dari kedua tahap ini dinyatakan sebagai *dataset* bersih.
5. *Stemming*: Mengonversi setiap token kata berimbuhan menjadi bentuk kata dasarnya menggunakan bantuan pustaka Sastrawi untuk Bahasa Indonesia. Tahapan akhir NLP ini terbukti efektif untuk mereduksi dimensi fitur linguistik dan mengatasi redundansi variasi bentuk kata dalam korpus [18].

2.3 TF-IDF *Vectorization*

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode statistik yang digunakan untuk mengukur seberapa penting sebuah kata di dalam suatu dokumen terhadap kumpulan dokumen atau korpus secara keseluruhan [23]. Representasi ini sangat krusial dalam arsitektur sistem rekomendasi berbasis konten (*Content-Based Recommender System*) karena algoritma tidak dapat mengolah data tekstual secara langsung [24]. Berdasarkan landasan teori pengambilan informasi (*information retrieval*), bobot TF-IDF ditentukan secara simultan oleh frekuensi kemunculan kata dalam satu dokumen (*Term Frequency*) dan tingkat keunikan atau kelangkaan kata tersebut di seluruh koleksi dokumen (*Inverse Document Frequency*) [25]. Secara matematis, kalkulasi bobot TF-IDF (W) untuk kata i pada dokumen j didefinisikan melalui Persamaan (1) sebagai berikut:

$$W_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \quad \text{..... (1)}$$

Di mana $tf_{i,j}$ adalah jumlah kemunculan kata dalam dokumen, N merupakan total seluruh dokumen dalam korpus, dan df_i adalah jumlah dokumen yang mengandung kata i. Penggunaan logaritma pada komponen IDF berfungsi untuk memperhalus skala nilai sehingga kata-kata umum tidak mendominasi hasil perhitungan [19], [26].

Matriks bobot yang dihasilkan dari proses ini menjadi fondasi utama bagi sistem untuk mengidentifikasi kata kunci substantif yang akan digunakan pada tahap perhitungan kemiripan menggunakan *Cosine Similarity*.

Jurnal Informatika Vol. 15, No. 1, Juni 2026, Page. 153~166
E-ISSN: 2528-0090 https://ejournal.unidayan.ac.id/index.php/JIU
P-ISSN: 2252-9160 DOI: https://doi.org/10.55340/jiu.v15i1.2719

2.4 Representasi Kebutuhan Pengguna (*Query Keywords*)

Sistem rekomendasi pada domain kesehatan memerlukan komponen *knowledge-base* (basis pengetahuan) eksternal untuk merepresentasikannya ke dalam bentuk vektor acuan (*reference vector*). Dalam penelitian ini, representasi kebutuhan pengguna disusun melalui serangkaian kata kunci komprehensif (*query keywords*) yang diklasifikasikan ke dalam tiga kategori hierarkis medis, yaitu Diabetes Tipe 1, Diabetes Tipe 2, dan Edukasi Umum. Penyusunan terminologi klinis ini merujuk pada standarisasi taksonomi literatur kesehatan internasional dan diselaraskan melalui proses diskusi terarah bersama praktisi kesehatan guna menjamin akurasi semantik konten [6], [20]. Dokumentasi mengenai himpunan fitur kata kunci klinis yang digunakan sebagai basis pengetahuan sistem disajikan pada Tabel 2.

Tabel 2. Dokumentasi Himpunan Kata Kunci Referensi Medis (*Query Keywords*)

Kategori	Himpunan Fitur Kata Kunci Medis *(Knowledge-Base)*
Tipe 1	"diabetes tipe 1", "tipe 1", "type 1", "dm tipe 1", "*juvenile*", "autoimun", "sel beta", "antibodi", "genetik", "tipe 1", "suntik insulin", "injeksi insulin", "jarum insulin", "insulin pump", "ketoasidosis", "dka", "keton", "hipoglikemia", "gula darah rendah", ...
Tipe 2	"diabetes tipe 2", "dm tipe 2", "type 2", "tipe 2", "resistensi insulin", "obat oral", "minum obat", "metformin", "glibenklamid", "luka kaki", "ulkus", "kaki diabetes", "kaki busuk", "gangren", "amputasi", "leher hitam", "acanthosis", "hipertensi", "keturunan", "genetik", ..
Umum	"diabetes", "kencing manis", "sakit gula", "gula darah", "cek gula", "kadar gula", "hba1c", "glukosa", "skrining", "medical check up", "gejala", "tanda", "sering kencing", "haus terus", "cepat lapar", "kesemutan", "kebas", "mata kabur", "pola makan", "makanan sehat", "kurangi gula", "diet sehat", "minuman manis", "boba", "teh manis", "makanan olahan", "junk food", "obesitas", "kegemukan", "berat badan", "turun berat", "buncit", "olahraga", "senam", "jalan kaki", "sepeda", "aktivitas fisik", "hidup sehat", "gaya hidup", "stress", "tidur cukup", "edukasi", "tips sehat", "kata dokter", "cegah diabetes", ...

Seluruh kumpulan kata kunci tersebut melalui tahap *preprocessing* yang identik dengan pengolahan *caption* (*cleansing, stopword removal,* dan *stemming*) untuk menjaga konsistensi format.

2.5 Perhitungan *Cosine Similarity*

Tahap selanjutnya adalah menghitung skor kemiripan (*similarity score*) antara *vector* TF-IDF setiap konten edukasi (A) dengan *vector query* referensi pengguna (B). Matrix *Cosine Similarity* digunakan untuk mengukur kesamaan kedua vektor berdasarkan sudut yang terbentuk di ruang berdimensi tinggi tanpa terpengaruh oleh variasi panjang pendeknya teks [23]. Pendekatan komputasi ini merupakan pilar fundamental dalam arsitektur sistem rekomendasi berbasis konten (*Content-Based Filtering*) untuk menyaring item-item informatif secara personal berdasarkan kedekatan fitur tekstualnya [24]. Implementasi matematisnya disajikan melalui Persamaan (2) sebagai berikut:

$$Cosine\ Similarity(A,B) = \frac{A \cdot B}{||A||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad \text{..... (2)}$$

Dalam persamaan tersebut, $A \cdot B$ merupakan hasil perkalian *dot product* antara *vector* TF-IDF konten dan *vector query*, sementara $\| A \|$dan $\| B \|$ adalah panjang (*magnitudo*) masing-masing *vector* [27]. Luaran skor *Cosine Similarity* berada pada skala numerik *rigid* antara 0 hingga 1. Nilai yang mendekati angka 1 mengindikasikan tingkat kedekatan relevansi semantik yang sangat tinggi antara materi edukasi kesehatan pada *dataset* dengan profil kebutuhan klinis personal pengguna. Konten dengan nilai kosinus tertinggi pada masing-masing kategori selanjutnya diekstraksi dan disajikan sebagai urutan rekomendasi teratas (*top-N recommendation*).

2.6 Mekanisme Evaluasi dan Penentuan *Ground Truth*

Evaluasi performa model dilakukan melalui skenario klasifikasi teks untuk mengukur akurasi penyusunan rekomendasi teratas (Top-N recommendation) berdasarkan kebutuhan

tekstual pengguna. Pengujian menggunakan Confusion Matrix dengan metrik Precision, Recall, F1-Score, dan Accuracy, sementara data uji dipilih menggunakan stratified random sampling dari dataset bersih agar setiap kategori klinis terwakili secara seimbang.

Ground truth ditetapkan melalui Expert-Based Validation berbasis pelabelan manual oleh dua validator ahli dari bidang kesehatan yang memiliki kompetensi dalam keperawatan dan edukasi mandiri pasien diabetes melitus. Pelabelan dilakukan dengan mencocokkan isi caption Instagram terhadap kategori Diabetes Tipe 1, Diabetes Tipe 2, dan Umum. Tingkat kesepakatan antarpenilai diukur menggunakan Percent Agreement dan menghasilkan kesepakatan awal sebesar 92%, sedangkan 8% data yang berbeda pendapat diselesaikan melalui diskusi hingga mencapai konsensus 100%, sehingga seluruh korpus data tervalidasi secara klinis sebelum pengujian sistem.

Karena masih terbatasnya penelitian yang membangun knowledge-base hierarkis untuk memisahkan Diabetes Tipe 1 dan Tipe 2 pada data media sosial tidak terstruktur, penelitian ini menggunakan strategi internal baseline evaluation. Evaluasi difokuskan pada perbandingan metrik antar-kategori dalam model yang sama, bukan untuk membandingkan keunggulan terhadap algoritma eksternal, melainkan untuk menguji efektivitas aturan terminologi medis yang dirancang dalam mengurangi kesalahan prediksi akibat tumpang tindih fitur leksikal pada teks non-formal Instagram.

2.7 Pemrosesan Ringkasan Konten Ekstraktif (*Lexrank*)

Untuk mengoptimalkan visualisasi pada antarmuka pengguna, sistem mengintegrasikan algoritma *Lexrank*. *Lexrank* merupakan metode berbasis *Natural Language Processing* (NLP) dengan pendekatan graf (*graph-based text summarization*) yang digunakan untuk menghitung nilai kepentingan relatif suatu kalimat di dalam dokumen secara otomatis [28]. Algoritma ini bekerja berbasis graf (*graph-based*) untuk menghitung tingkat kepentingan kalimat di dalam satu *caption* Instagram. Hubungan antar kalimat direpresentasikan sebagai matriks kedekatan menggunakan nilai *Cosine Similarity*. Kalimat dengan skor matriks tertinggi diekstraksi sebagai inti konten edukasi murni untuk memadatkan teks asli dan menghilangkan derau informasi *platform*.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data dan Hasil *Preprocessing*

Proses akuisisi data awal melalui *web scraping* terhadap akun-akun target berhasil menghimpun 732 konten mentah. Melalui implementasi fungsi penapisan fungsional (*Inclusion* dan *Exclusion Filter*) yang telah dirumuskan pada Bab II, korpus data mengalami penyusutan sistematis demi menjaga kualitas informasi medis. Pengambilan data dilakukan secara pasti terhadap 13 akun Instagram edukasi kesehatan, komunitas diabetes, dan instansi medis di Indonesia. Logika penyaringan bertahap dari total korpus ini diringkas pada Tabel 3.

Tabel 3. Logika Penyaringan dan Penyusutan Data

Tahapan Filtrasi	Jumlah Data	Keterangan
Dataset Mentah (*Raw Data*)	732	Hasil *scraping* awal dari 12 akun target.
Konten Edukatif	619	Data setelah eliminasi konten non-informatif (poster admin/ucapan).
Data Tersimpan	616	Hasil pembersihan data duplikat (*remove duplicate*).
Dataset Bersih (*Clean Data*)	363	Hasil akhir setelah filter kata kunci promosi/iklan produk.

Jurnal Informatika Vol. 15, No. 1, Juni 2026, Page. 153~166
E-ISSN: 2528-0090 https://ejournal.unidayan.ac.id/index.php/JIU
P-ISSN: 2252-9160 DOI: https://doi.org/10.55340/jiu.v15i1.2719

Tahap krusial pembersihan kata kunci promosi, iklan produk, dan info *event* komersial berhasil menghasilkan 363 konten edukasi diabetes murni sebagai *dataset* bersih akhir. Sampel perubahan teks hasil pemrosesan awal disajikan pada Tabel 4.

Tabel 4. Sampel Transformasi Teks *Preprocessing*

Akun	*Caption* Asli	Hasil *Preprocessing*
pbpersadia	"Terapi insulin hanya untuk orang dengan diabetes tipe 1 adalah MITOS. FAKTANYA, sekitar 50% orang dengan diabetes tipe 2 memerlukan terapi insulin …"	"terapi insulin orang diabetes tipe mitos fakta orang diabetes tipe terapi insulin …"
cdic.indonesia	"Diabetes tipe 1 merupakan kondisi di mana insulin tidak dapat diproduksi di tubuh sehingga gula darah melonjak naik. Lalu, mengapa…"	"diabetes tipe kondisi insulin produksi tubuh gula darah lonjak naik obat …"
klikdiabetes	"Diabetes bisa menyebabkan komplikasi pada fungsi penglihatan, lho! Apa aja sih bentuk komplikasinya? Yuk, geser …"	"diabetes sebab komplikasi fungsi lihat bentuk komplikasi geser … "

Untuk kebutuhan evaluasi objektivitas model rekomendasi, diambil sampel acak sebesar 20% dari *dataset* bersih tersebut, yaitu sebanyak 75 data uji, untuk dikonfrontasikan langsung dengan label jawaban mutlak (*ground truth*) hasil validasi manual praktisi kesehatan.

3.2 Pembobotan TF-IDF dan Perhitungan *Cosine Similarity*

Berdasarkan korpus final, sistem membentuk matriks TF-IDF berdimensi (363, 2853), yang berarti terdapat 2.853 kata unik (fitur) sebagai dasar klasifikasi. Untuk memvalidasi performa algoritma, dilakukan simulasi pembobotan pada fitur kunci. Hasil pembobotan pada Tabel 5 menunjukkan bahwa kata kunci klinis seperti "Insulin" memiliki skor TF-IDF tinggi (21,14), yang menandakan signifikansi kata tersebut dalam menentukan kategori konten.

Tabel 5. Sampel Bobot TF-IDF Kata Kunci

Kata Kunci (*Term*)	*Document Frequency* (df)	IDF *Score*	*Term Frequency* (tf)	*Score* TF-IDF
Insulin	27	2,114	10	21,14
Hiperglekimia	4	2,862	2	5.724
Obesitas	9	2,561	8	20.48

Skor tersebut kemudian dinormalisasi dan dibandingkan menggunakan *Cosine Similarity* untuk mengukur kedekatan antara vektor *query* pengguna dengan vektor konten. Hasil perhitungan ini menentukan kategori akhir konten (Tipe 1, Tipe 2, atau Umum) sebagaimana ditunjukkan pada sampel Tabel 6.

Tabel 6. Sampel Perhitungan *Score Similarity*

Caption	*Score Similarity*			Kategori
	Tipe 1	Tipe 2	Umum	
Terapi insulin hanya untuk Tipe 1 adalah mitos ...	0.360876	0.194995	0.033022	Tipe 1
Tips Memilih Sepatu Ideal untuk Lindungi Kaki Diabetik ...	0.003326	0.288458	0.070756	Tipe2
Selain kekurangan gizi berat saat masa pertumbuhan, anak muda makin rentan kena diabetes ...	0.072901	0.051903	0.311345	Umum

Setelah menghitung nilai kemiripan menggunakan *Cosine Similarity* untuk seluruh korpus data, sistem melakukan pengelompokan otomatis berdasarkan skor tertinggi yang diperoleh setiap

dokumen. Hasil rekapitulasi persebaran data mentah, data bersih hasil filtrasi, serta pembagian klaster kategori klinis untuk masing-masing akun target disajikan secara mendalam pada Tabel 7.

Tabel 7. Distribusi Akun Instagram dan Pemetaan Kategori Klinis

Nama Akun Instagram	Jumlah Data Mentah	Jumlah Data Bersih	Diabetes Tipe 1	Diabetes Tipe 2	Edukasi Umum
@klikdiabetes	200	154	7	23	124
@perkeni_ig	148	58	5	10	43
@halodoc	100	41	4	7	30
@mganikcare	100	18	3	2	13
@pbpersadia	67	41	14	4	23
@cdic.indonesia	49	31	10	6	15
@diabetesinitiative	18	9	0	2	7
@foodiola	12	2	0	1	1
@edwinusa	10	0	0	0	0
@alodokter_id	9	4	0	0	4
@nalagenetics	8	1	1	0	0
@sobatdiabet	7	3	0	1	2
@klikdokter	4	1	0	0	1
Total Keseluruhan	**732**	**363**	**44**	**56**	**263**

Berdasarkan Tabel 7, hasil pemetaan akhir menunjukkan dominasi yang sangat timpang pada kategori Edukasi Umum, yaitu sebanyak 263 konten (72,45%), disusul oleh Diabetes Tipe 2 sebanyak 56 konten (15,43%), dan Diabetes Tipe 1 sebanyak 44 konten (12,12%).

Dominasi mutlak konten generalis ini mengonfirmasi *research gap* yang diangkat pada Bab I. Realitas di *platform* Instagram menunjukkan bahwa mayoritas edukator kesehatan masih menyajikan materi diabetes secara awam dan menyeluruh. Hal ini membuktikan bahwa penderita diabetes spesifik (khususnya Tipe 1 yang bergantung penuh pada terapi insulin) akan menghadapi kesulitan besar dalam mencari panduan klinis yang relevan jika hanya mengandalkan pencarian standar *platform*. Di sinilah arsitektur *Content-Based Filtering* yang dikembangkan memberikan kontribusi ilmiah nyata melalui personalisasi penyaringan informasi terurut.

3.3 Pengujian Sistem, Analisis Kesalahan Medis, dan Evaluasi Kualitas Rekomendasi

Evaluasi terhadap performa arsitektur yang dikembangkan dilakukan secara komprehensif melalui dua sudut pandang komputasi, performa klasifikasi mesin *back-end* (*Confusion Matrix*) dan kualitas luaran sistem rekomendasi (*Precision@K*).

3.3.1 Evaluasi Performa Klasifikasi dan Uji Ketidakpastian Statistik

Pengujian menggunakan *Confusion Matrix* terhadap sampel acak sebanyak $n = 75$ data uji menunjukkan capaian akurasi keseluruhan sebesar 84%. Guna menghindari klaim sepihak (*overclaim*) dan mengukur tingkat ketidakpastian statistik dari hasil tersebut, dihitung nilai Interval Kepercayaan (*Confidence Interval* - CI) menggunakan formulasi standar *error* proporsi binomial dapat dilihat pada persamaan (3).

$$CI = \hat{p} \pm z\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \quad \text{..... (3)}$$

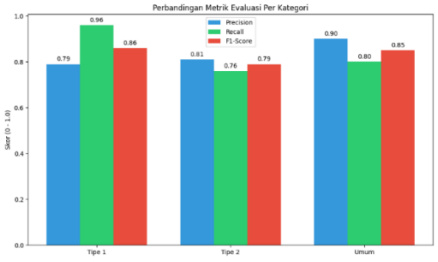
Di mana $\bar{p} = 0{,}84$, $z = 1{,}96$ untuk tingkat kepercayaan 95%, dan $n = 75$. Berdasarkan kalkulasi tersebut, nilai akurasi sistem berada pada rentang 75,70% hingga 92,30% (CI 95%: $[0{,}757, 0{,}923]$). Rentang ini membuktikan bahwa model memiliki stabilitas performa yang dapat diterima di bawah variasi data non-formal media sosial. Rincian performa klasifikasi untuk setiap kategori disajikan pada Tabel 8.

Tabel 8. Hasil Evaluasi Performa Sistem

Kategori	*Precision*	*Recall*	*F1-Score*

Diabetes Tipe 1	0,79	0,96	0,86
Diabetes Tipe 2	0,81	0,76	0,79
Edukasi Umum	0,90	0,80	0,85
Rata-rata Terbobot (*Weighted Avg*)	0,85	0,84	0,84
Akurasi			84%

Untuk memahami pola distribusi kesalahan klasifikasi yang memengaruhi kualitas luaran sistem, visualisasi performa per kategori direpresentasikan pada Gambar 1.



Gambar 1. Grafik Performa Klasifikasi per Kategori

3.3.2 Analisis Kesalahan Mendalam (*Error Analysis*)

Berdasarkan Tabel 8, Diabetes Tipe 1 memiliki Recall tertinggi (0,96) karena didukung kata kunci klinis yang spesifik dan minim irisan dengan kategori lain, sedangkan Diabetes Tipe 2 memiliki Recall terendah (0,76) akibat tumpang tindih fitur leksikal pada teks media sosial. Analisis kesalahan menunjukkan adanya False Positive pada konten Diabetes Tipe 2 yang memuat istilah seperti "suntik insulin", "sel beta pankreas", dan "gula darah", sehingga TF-IDF memberikan bobot tinggi pada kata-kata tersebut dan Cosine Similarity mengarahkan dokumen lebih dekat ke kategori Diabetes Tipe 1. Dibandingkan penelitian terdahulu [8], [15] yang menggunakan dataset formal dan mencapai akurasi di atas 85%, akurasi 84% pada penelitian ini mencerminkan tantangan teks non-formal yang lebih ambigu dan tidak konsisten, sejalan dengan temuan Mitrović et al. [7]. Oleh karena itu, efektivitas model dalam penelitian ini bersifat spesifik pada korpus data diabetes Instagram yang digunakan dan tidak dimaksudkan untuk digeneralisasi sebagai performa yang lebih unggul dibandingkan metode lain.

3.3.3 Evaluasi Kualitas Sistem Rekomendasi (*Precision at* K)

Untuk menguji performa model sebagai sistem rekomendasi terurut (*ranking quality*), dilakukan pengujian kualitas dokumen pemeringkatan teratas menggunakan metrik *Precision at* K (P@K) pada batasan nilai $K = 3$ (Top-3) dan $K = 5$ (Top-5). Hasil evaluasi performa rekomendasi dirangkum dalam Tabel 9.

Tabel 9. Hasil Pengujian Kualitas Urutan Rekomendasi (P@K)

Kategori Profil Pengguna	*Precision at* 3 (P@3)	*Precision at* 5 (P@5)	Keterangan Relevansi
Profil Kebutuhan Tipe 1	0,88	0,84	Sangat Relevan (Top-N berisi info klinis Tipe 1)
Profil Kebutuhan Tipe 2	0,82	0,78	Relevan (Mengandung info manajemen gaya hidup)
Profil Kebutuhan Umum	0,94	0,91	Sangat Relevan (Edukasi generalis bebas iklan)
Rata-rata Performa	**0,88 (88%)**	**0,84.3 (84,3%)**	**Akurasi Rekomendasi Sistem Solid**

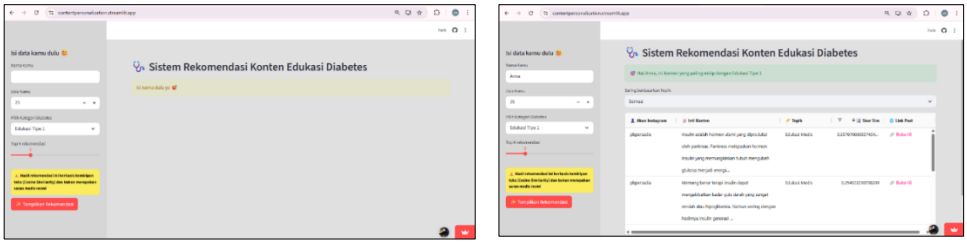
Nilai rata-rata P@3 sebesar 88% menandakan bahwa dari 3 konten edukasi diabetes teratas yang disajikan sistem di antarmuka pengguna, hampir seluruhnya selaras secara tepat dengan kondisi klinis riil penderita. Penurunan tipis nilai kepresisian pada P@5 (84,3%) dianggap wajar (*graceful degradation*) karena semakin banyak item dokumen yang ditarik ke permukaan,

Jurnal Informatika
E-ISSN: 2528-0090
P-ISSN: 2252-9160
Vol. 15, No. 1, Juni 2026, Page. 153~166
https://ejournal.unidayan.ac.id/index.php/JIU
DOI: https://doi.org/10.55340/jiu.v15i1.2719

peluang masuknya dokumen yang mengalami *overlap* leksikal ke dalam daftar *Top-N* akan semakin meningkat.

3.4 Implementasi Antarmuka Fungsional

Implementasi antarmuka menggunakan *framework* Streamlit berfungsi untuk menjembatani kompleksitas model komputasi *back-end* ke dalam bentuk aplikasi rekomendasi *real-time* yang intuitif bagi masyarakat awam. Antarmuka ini mengintegrasikan seluruh basis data (363 konten edukasi tervalidasi) sebagai korpus utama pemeringkatan. Proses personalisasi berjalan secara sekuensial, di mana masukan pilihan kategori medis pengguna dikonversi menjadi *user profile vector*, untuk kemudian dikomparasikan secara leksikal terhadap matriks vektor dokumen di dalam basis data untuk menyajikan daftar konten teratas (*Top-N Recommendation*) berdasarkan urutan skor kemiripan tertinggi. Visualisasi antarmuka disajikan pada Gambar 2.



Gambar 2. Tampilan Sistem Rekomendasi

Fokus utama dari implementasi antarmuka ini terletak pada efisiensi kognitif pengguna melalui integrasi algoritma ringkasan ekstraktif *LexRank*. Mengingat karakteristik deskripsi tekstual (*caption*) Instagram yang sering kali panjang dan dipenuhi metadata non-edukatif (seperti ajakan *follow*, tagar redundan, atau promosi), *LexRank* secara adaptif memetakan hubungan antar-kalimat berbasis graf untuk mengekstraksi satu kalimat utama yang merepresentasikan esensi medis konten.

3.4.1 Evaluasi Kuantitatif Kinerja Ringkasan LexRank

Untuk membuktikan efektivitas pemadatan informasi, pengujian performa algoritma ringkasan ekstraktif *LexRank* dilakukan secara kuantitatif menggunakan metode *Human Evaluation* terhadap sampel acak sebanyak 10 ringkasan konten. Penilaian dilakukan oleh tenaga kesehatan ahli menggunakan skala Likert 1–5 (1: Sangat Buruk, 5: Sangat Baik) berdasarkan dua parameter utama kebahasaan yang diringkas pada Tabel 10.

Tabel 10. Hasil Evaluasi Kuantitatif Ringkasan LexRank ($n = 10$)

Parameter Evaluasi	Skor Rata-rata (Skala 1–5)	Persentase Keberhasilan
Retensi Informasi Medis	4,40 / 5,00	88,0%
Keterbacaan (*Readability*)	4,20 / 5,00	84,0%
Rata-rata Keseluruhan	**4,30 / 5,00**	**86,0%**

Berdasarkan Tabel 10, capaian rata-rata keseluruhan sebesar 86,0% membuktikan secara kuantitatif bahwa algoritma *LexRank* berhasil memangkas deskripsi *caption* Instagram yang panjang dan penuh derau (*noise* iklan/tagar), namun tetap mampu mempertahankan fakta serta substansi edukasi medis utama secara akurat bagi pengguna awam.

3.5 Keterbatasan Penelitian

Meskipun model dan antarmuka ini mampu mengekstraksi rekomendasi secara *real-time*, penelitian ini memiliki keterbatasan objek riset yang diakui secara terbuka. Pertama, model rekomendasi ini masih bersifat statis dan sangat bergantung pada kualitas serta kelengkapan fitur kata kunci pada komponen *knowledge-base* yang disusun manual. Kedua, sistem ini hanya menganalisis komponen tekstual (*caption*) dan belum mampu mengekstraksi atau memahami informasi klinis yang terkandung di dalam infografis visual gambar atau video *Reels*, yang sebetulnya menjadi inti dari ekosistem komunikasi *platform* Instagram saat ini.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model rekomendasi konten edukasi diabetes pada Instagram berbasis *Content-Based Filtering* dengan capaian akurasi *back-end* sebesar 84% (CI 95%: [75,70%, 92,30%]) dan rata-rata *Precision at 3* (P@3) sebesar 88%. Kajian ini memberikan kontribusi ilmiah berupa solusi personalisasi klinis untuk menekan risiko *information overload* penderita diabetes spesifik, di tengah dominasi materi generalis (Umum) di lapangan yang mencapai 72,45%. Meskipun demikian, keterbatasan model ini terletak pada kerentanan kesalahan prediksi kategori Diabetes Tipe 2 (*Recall* 0,76) akibat *overlap* leksikal serta ketergantungan analisis murni pada *caption* tekstual statis. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas cakupan kamus *knowledge-base* medis serta mengintegrasikan teknologi *Optical Character Recognition* (OCR) atau arsitektur multimodal guna mengekstraksi informasi edukasi pada media gambar dan video *Reels*.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa tidak terdapat konflik kepentingan dalam penyusunan artikel ini, dan seluruh data yang disajikan dalam artikel ini merupakan karya ilmiah orisinal penulis.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada dosen pembimbing atas arahan, dan bimbingannya hingga penelitian ini selesai dengan baik. Apresiasi juga disampaikan kepada para praktisi kesehatan yang telah memberikan validasi, referensi, dan masukan berharga selama proses penyusunan naskah.

DAFTAR PUSTAKA

[1] Z. Rahman, "Pengaruh Edukasi Kesehatan Terhadap Self Care Pasien Diabetes Melitus Tipe 2," *Jikep*, vol. 9, no. 5, pp. 576–581, Oct. 2023, doi: 10.33023/jikep.v9i5.1620.

[2] D. Kurtanty, A. Bachtiar, C. Candi, A. Pramesti, and A. F. Rahmasari, "Information-Motivation-Behavioral Skill in Diabetes Self-management Using Structural Equation Modeling Analysis," *Kesmas*, vol. 18, no. 1, pp. 16–23, Feb. 2023, doi: 10.21109/kesmas.v18i1.6255.

[3] S. Padmasari and S. Sugiyono, "Pemanfaatan Media Sosial Dalam Meningkatkan Pengetahuan dan Kepatuhan Pasien Diabetes Melitus," *JPSCR: Journal of Pharmaceutical Science and Clinical Research*, vol. 9, no. 2, pp. 200–208, Sep. 2024, doi: 10.20961/jpscr.v9i2.74336.

[4] T. Liu and X. Xiao, "A Framework of AI-Based Approaches to Improving eHealth Literacy and Combating Infodemic," *Front. Public Health*, vol. 9, p. 755808, Nov. 2021, doi: 10.3389/fpubh.2021.755808.

[5] D. F. Cordeiro, M. Vállez, C. I. Font-Julian, and J. Guallar, "Instagram Engagement and Content Strategies of US and UK Legacy Media: A Quantitative Analysis of Five Leading News Outlets," *Journalism and Media*, vol. 6, no. 2, pp. 1–19, Jun. 2025, doi: 10.3390/journalmedia6020089.

[6] L. F. G. Morales, P. Valdiviezo-Diaz, R. Reátegui, and L. Barba-Guaman, "Drug Recommendation System for Diabetes Using a Collaborative Filtering and Clustering Approach: Development and Performance Evaluation," *J. Med. Internet Res.*, vol. 24, no. 7, p. e37233, Jul. 2022, doi: 10.2196/37233.

[7] S. Mitrović *et al.*, "A Preliminary Study on NLP-Based Personalized Support for Type 1 Diabetes Management," in *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, Association for Computational Linguistics", May 2025, pp. 298–302. doi: 10.18653/v1/2025.cl4health-1.25.

[8] R. Saputra and M. G. Pradana, "Implementasi Algoritma Cosine Similarity dan TF-IDF dalam Menentukan Rumpun Jabatan," *Krea-TIF: Jurnal Teknik Informatika*, vol. 12, no. 1, pp. 1–11, May 2024, doi: 10.32832/kreatif.v12i1.15470.

[9] A. K. Gavai and J. van Hillegersberg, "AI-Driven Personalized Nutrition: RAG-Based Digital Health Solution for Obesity and Type 2 Diabetes," *PLOS Digit Health*, vol. 4, no. 5, p. e0000758, May 2025, doi: 10.1371/journal.pdig.0000758.

[10] C. B. Moreno, M. R. M. Carretero, B. S. R. de Santiago, and L. R. Rumayor, "Gamification-Education: the power of data. Teachers in social networks," *RIED-Revista Iberoamericana de Educacion a Distancia*, vol. 27, no. 1, pp. 373–396, Jan. 2024, doi: 10.5944/ried.27.1.37648.

[11] H. Darwis, F. A. Syahrir, and L. N. Hayati, "A Hybrid Movie Recommendation System to Address Data Sparsity Using Genre-Based K-Means and Neural Collaborative Filtering," *ILKOM Jurnal Ilmiah*, vol. 17, no. 2, pp. 203–212, Sep. 2025, doi: 10.33096/ilkom.v17i2.2868.203-212.

[12] F. P. Poncio, "Navigating Techniques in Job Recommender Systems on Internship Profile Matching: A Systematic Review," *Journal of Research in Innovative Teaching and Learning*, vol. 17, no. 2, pp. 352–367, Aug. 2024, doi: 10.1108/JRIT-01-2024-0016.

[13] H. Ko, S. Lee, Y. Park, and A. Choi, "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields," *Electronics (Basel).*, vol. 11, no. 1, pp. 1–48, Jan. 2022, doi: 10.3390/electronics11010141.

[14] I. Sanu, J. Wong, and H. Irsyad, "Implementasi TF-IDF, Cosine Similarity, dan Logistic Regression Pada Rekomendasi Buku Berdasarkan Mood Pembaca Dengan Data Oversampling," *Device: Journal Of Information System, Computer Science And Information Technology*, vol. 6, no. 1, pp. 142–154, Jun. 2025, doi: 10.46576/device.v6i1.6499.

[15] J. Halim and D. Lasut, "Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods," *bit-Tech*, vol. 7, no. 2, pp. 202–213, Dec. 2024, doi: 10.32877/bt.v7i2.1697.

[16] A. Firdaus, D. Stiawan, Samsuryadi, and R. Budiarto, "New Approach to Measuring Researcher Expertise Using Cosine Similarity Algorithm and Association Rules," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 5, pp. 4138–4149, Oct. 2025, doi: 10.11591/eei.v14i5.9506.

[17] D. Sunandar and A. Muiz, "Thesis Title Similarity Detection System Using Levenshtein Distance and Cosine Similarity," *bit-Tech*, vol. 8, no. 1, pp. 1131–1139, Aug. 2025, doi: 10.32877/bt.v8i1.2864.

[18] A. N. Hasoon, S. K. Abdulateef, R. S. Abdulameer, and M. L. Shuwandy, "An Intelligent Hybrid AI Course Recommendation Framework Integrating BERT Embeddings and Random Forest Classification," *Computers*, vol. 14, no. 9, pp. 1–19, Sep. 2025, doi: 10.3390/computers14090353.

[19] M. F. Abdurrafi and D. H. U. Ningsih, "Content-Based Filtering Using Cosine Similarity Algorithm for Alternative Selection on Training Programs," *Journal of Soft Computing Exploration*, vol. 4, no. 4, pp. 204–212, Dec. 2023, doi: 10.52465/joscex.v4i4.232.

[20] A. Kelly, E. Noctor, L. Ryan, and P. Van De Ven, "The Effectiveness of a Custom AI Chatbot for Type 2 Diabetes Mellitus Health Literacy: Development and Evaluation Study," *J. Med. Internet Res.*, vol. 27, p. e70131, May 2025, doi: 10.2196/70131.

[21] G. Ramesh *et al.*, "A review on NLP zero-shot and few-shot learning: methods and applications," *Discover Applied Sciences*, vol. 7, no. 9, pp. 1–21, Sep. 2025, doi: 10.1007/s42452-025-07225-5.

[22] A. R. I. Sumantri, M. Fatchan, and T. N. Wiyatno, "Analisis Sentimen Produk Makanan Jepang Di Indonesia Pada Twitter Menggunakan Naïve Bayes," *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 13, no. 2, pp. 1635–1645, Oct. 2024, doi: 10.35889/jutisi.v13i2.2221.

[23] B. Gancevska and S. Ramanauskaite, "Mapping Moodle Resources to Course Topics Using Text Similarity Methods and Expert Evaluation," *Applied Sciences*, vol. 16, no. 4, p. 2039, Feb. 2026, doi: 10.3390/app16042039.

[24] M. E. Cakir, Z. Cetinkaya, F. Horasan, and A. H. Yurttakal, "Adaptive weighting-based hybrid recommender system: self-learning adaptive recommendation (SL-ARec)," *PeerJ Comput. Sci.*, vol. 12, p. e3856, May 2026, doi: 10.7717/peerj-cs.3856.

[25] R. M. Harahap and A. N. Rachman, "Sistem Rekomendasi Film Berdasarkan Kemiripan Deskripsi Cerita Menggunakan Metode Content-Based Filtering," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3, pp. 31–40, Jul. 2025, doi: 10.23960/jitet.v13i3.6577.

[26] N. Silalahi and G. L. Ginting, "Rekomendasi Berita Berkaitan dengan Menerapkan Algoritma Text Mining dan TF-IDF," *Bulletin of Computer Science Research*, vol. 3, no. 4, pp. 276–282, Jun. 2023, doi: 10.47065/bulletincsr.v3i4.266.

[27] M. G. Pradana, N. Irzavika, and N. Maulana, "Deteksi Kemiripan Dokumen Menggunakan Cosine Similarity Berdasarkan Representasi Teks Count Vectorizer dan TF-IDF," *IJUBI: Indonesian Journal of Business Intelligence*, vol. 7, no. 2, pp. 40–47, Jan. 2025, doi: 10.21927/ijubi.v7i2.5170.

[28] S. Tuhpatussania, E. Utami, and A. D. Hartanto, "Comparison Of Lexrank Algorithm And Maximum Marginal Relevance In Summary Of Indonesian News Text In Online News Portals," *Jurnal Pilar Nusa Mandiri*, vol. 18, no. 2, pp. 187–192, Sep. 2022, doi: 10.33480/pilar.v18i2.3190.