

KNOWLEDGE DISCOVERY TERHADAP INFORMASI KESEHATAN DI MEDIA X MENGGUNAKAN TOPIC MODELING BERBASIS NON-NEGATIVE MATRIX FACTORIZATION (NMF)

Nabila Ramadhani, Dinda Aulika Elfarin Aritonang, Ken Ditha Tania, Ahmad Rifai

Sistem Informasi, Universitas Sriwijaya

Jl. Raya Palembang-Prabumulih, KM 32 Indralaya, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia

nabila.r086@gmail.com

ABSTRAK

Media sosial menjadi salah satu sumber utama penyebaran dan pertukaran informasi kesehatan di masyarakat. Media X (Twitter) memungkinkan pengguna untuk berbagi pengalaman, pertanyaan, dan opini terkait berbagai isu kesehatan secara real-time. Namun, informasi kesehatan yang tersebar di media sosial bersifat tidak terstruktur dan jumlahnya sangat besar sehingga sulit dianalisis secara manual untuk mengetahui kebutuhan informasi kesehatan masyarakat. Oleh karena itu, penelitian ini bertujuan untuk melakukan *knowledge discovery* terhadap informasi kesehatan yang dibagikan di media sosial X menggunakan pendekatan *topic modeling* berbasis Non-Negative Matrix Factorization (NMF). Data penelitian diperoleh melalui proses *web crawling* menggunakan kata kunci terkait kesehatan pada periode November 2025 hingga Februari 2026 dan menghasilkan 1.338 tweet setelah melalui tahap pembersihan data. Tahapan penelitian meliputi *data preprocessing*, pembobotan Term Frequency-Inverse Document Frequency (TF-IDF), serta pemodelan topik menggunakan metode NMF. Hasil penelitian menunjukkan bahwa model NMF mampu mengidentifikasi lima topik utama, yaitu kesehatan mental, *self-diagnosis* dan konsultasi profesional, edukasi kesehatan dan layanan medis, mitos kesehatan, serta biaya dan akses pengobatan. Analisis distribusi topik menunjukkan bahwa kesehatan mental merupakan topik yang paling dominan dibahas oleh pengguna media sosial.

Kata kunci : *Knowledge Discovery, Machine Learning, Informasi Kesehatan, Topic Modeling, Non-Negative Matrix Factorization*

1. PENDAHULUAN

Informasi kesehatan merupakan salah satu jenis informasi yang paling sering dicari dan dibutuhkan oleh masyarakat dalam kehidupan sehari-hari [1].

Masyarakat tidak hanya bergantung pada tenaga medis atau fasilitas kesehatan formal, tetapi juga aktif mencari, berbagi, dan mendiskusikan berbagai isu kesehatan melalui media sosial [2].

Media X menjadi salah satu media yang banyak dimanfaatkan untuk menyampaikan pengalaman pribadi, keluhan kesehatan, pertanyaan, serta opini terkait kondisi kesehatan tertentu secara terbuka dan real-time. Aktivitas tersebut menghasilkan kumpulan data teks yang mencerminkan kebutuhan, perhatian, dan persepsi masyarakat terhadap berbagai isu kesehatan [3].

Informasi kesehatan yang tersebar di Media X bersifat tidak terstruktur, beragam, dan jumlahnya terus bertambah. Kondisi ini menjadikan data media sosial sebagai sumber informasi yang kaya, namun sulit untuk dianalisis secara manual. Di balik banyaknya unggahan tersebut, tersimpan pola-pola tertentu yang dapat memberikan gambaran mengenai topik kesehatan yang paling sering dibahas, isu yang sedang menjadi perhatian publik, serta kebutuhan informasi kesehatan yang belum sepenuhnya terpenuhi. Oleh karena itu, diperlukan pendekatan analitis yang mampu mengolah data teks dalam jumlah besar secara sistematis dan objektif.

Salah satu pendekatan yang dapat digunakan untuk menggali informasi bermakna dari data teks

dalam jumlah besar adalah *knowledge discovery*, yaitu proses untuk mengekstraksi pengetahuan baru, pola, dan hubungan tersembunyi dari kumpulan data yang besar dan kompleks [4].

Dalam konteks data teks, proses ini memungkinkan data mentah berupa teks diolah menjadi informasi yang memiliki nilai analitis dan dapat digunakan sebagai dasar dalam memahami fenomena tertentu.

Dalam proses *knowledge discovery* pada data teks, salah satu teknik yang sering digunakan untuk menganalisis pola topik dalam kumpulan dokumen adalah *topic modeling*, karena mampu mengidentifikasi topik-topik utama yang muncul dalam sekumpulan dokumen berdasarkan pola kemunculan kata [5].

Salah satu metode yang banyak digunakan dalam teknik *topic modeling* adalah Non-Negative Matrix Factorization (NMF). Metode ini bekerja dengan memfaktorkan matriks dokumen-kata menjadi matriks dokumen-topik dan matriks topik-kata sehingga menghasilkan topik yang lebih koheren dan mudah diinterpretasikan. Karakteristik tersebut menjadikan NMF sesuai untuk menganalisis data teks pendek dan informal seperti unggahan pada Media X.

Berdasarkan kebutuhan untuk menganalisis pola informasi kesehatan pada data media sosial tersebut, penelitian ini bertujuan untuk menerapkan metode Non-Negative Matrix Factorization (NMF) dalam proses *knowledge discovery* guna mengidentifikasi pola topik dan diskursus informasi kesehatan yang

muncul pada percakapan pengguna media sosial X. Penelitian ini diharapkan dapat mengidentifikasi topik-topik utama informasi kesehatan yang dibahas oleh masyarakat serta menghasilkan wawasan yang dapat dimanfaatkan untuk memahami kebutuhan informasi kesehatan di ruang digital. Hasil penelitian ini juga diharapkan dapat menjadi referensi bagi tenaga kesehatan, pengelola media digital, maupun pihak terkait dalam merancang strategi komunikasi dan edukasi kesehatan yang lebih efektif.

2. TINJAUAN PUSTAKA

Perkembangan teknologi informasi dan media sosial telah menghasilkan jumlah data teks yang sangat besar dan terus meningkat setiap harinya. Data tersebut umumnya bersifat tidak terstruktur sehingga memerlukan teknik khusus untuk mengekstraksi informasi yang terkandung di dalamnya. Salah satu pendekatan yang banyak digunakan adalah *text mining*, yaitu proses untuk menemukan pola atau pengetahuan yang tersembunyi dari kumpulan data teks dalam jumlah besar. Melalui teknik ini, informasi yang awalnya tidak terstruktur dapat diolah menjadi pengetahuan yang lebih terorganisasi dan mudah dianalisis [6].

Dalam proses penemuan pengetahuan dari data, dikenal konsep *Knowledge Discovery in Databases* (KDD) yang terdiri dari beberapa tahapan penting, seperti seleksi data, *preprocessing*, transformasi data, pemodelan, serta evaluasi hasil. Tahapan *preprocessing* memiliki peranan penting dalam analisis data teks karena data dari media sosial biasanya mengandung banyak *noise* seperti singkatan, simbol, maupun kesalahan penulisan. Proses pembersihan data yang optimal akan sangat mempengaruhi kualitas model yang dihasilkan dalam analisis teks [7].

Media sosial saat ini juga menjadi salah satu sumber informasi yang penting dalam berbagai bidang, termasuk kesehatan. Data percakapan yang dihasilkan dari media sosial dapat dianalisis melalui pendekatan *knowledge discovery* untuk memperoleh wawasan mengenai persepsi dan kebutuhan informasi masyarakat terhadap suatu isu tertentu [8].

Banyak masyarakat yang menggunakan media sosial untuk berbagi pengalaman, opini, maupun informasi terkait kondisi kesehatan dan layanan medis. Hal ini menyebabkan media sosial menjadi sumber data yang potensial untuk memahami persepsi masyarakat terhadap isu kesehatan.

Penelitian oleh [9] menunjukkan bahwa analisis percakapan masyarakat di Twitter mengenai isu stunting dapat memberikan gambaran mengenai topik-topik utama yang sedang dibahas oleh masyarakat, seperti gizi anak, program pemerintah, serta upaya pencegahan stunting.

Untuk menganalisis kumpulan data teks dalam jumlah besar, salah satu metode yang banyak digunakan adalah *topic modeling*. *Topic modeling* merupakan teknik analisis teks yang bertujuan untuk

mengidentifikasi topik-topik utama yang terdapat dalam sekumpulan dokumen secara otomatis. Metode ini memungkinkan peneliti untuk mengelompokkan dokumen berdasarkan kemunculan kata yang memiliki keterkaitan sehingga dapat diketahui tema-tema utama yang dibahas dalam suatu kumpulan teks [10].

Salah satu metode *topic modeling* yang paling banyak digunakan dalam penelitian adalah Latent Dirichlet Allocation (LDA).

Beberapa penelitian sebelumnya telah menerapkan metode LDA dalam analisis data media sosial.

Penelitian yang dilakukan oleh [11] menggunakan metode LDA untuk menganalisis cuitan masyarakat terkait penyakit tropis di Indonesia. Hasil penelitian menunjukkan bahwa metode tersebut mampu mengidentifikasi beberapa topik utama yang berkaitan dengan isu kesehatan seperti malaria, demam berdarah, dan upaya pencegahannya. Hal ini menunjukkan bahwa *topic modeling* dapat digunakan untuk memahami pola diskursus kesehatan di media sosial.

Selain LDA, metode lain yang juga banyak digunakan dalam *topic modeling* adalah Non-Negative Matrix Factorization (NMF). Metode NMF bekerja dengan melakukan dekomposisi matriks dokumen menjadi dua matriks non-negatif yang merepresentasikan hubungan antara kata dan topik. Pendekatan ini memungkinkan terbentuknya representasi topik yang lebih mudah diinterpretasikan karena setiap topik direpresentasikan oleh kumpulan kata dengan bobot tertentu [12].

Penelitian oleh [13] menunjukkan bahwa pendekatan *topic mining* berbasis NMF juga dapat digunakan dalam proses *knowledge discovery* untuk mengidentifikasi kebutuhan informasi kesehatan pengguna media sosial secara lebih sistematis.

Beberapa penelitian menunjukkan bahwa pendekatan *topic modeling* juga dapat diterapkan pada data media sosial untuk memahami opini publik.

Penelitian oleh [12] menggunakan metode LDA untuk menganalisis percakapan masyarakat di Twitter terkait kepercayaan terhadap institusi publik. Penelitian tersebut menunjukkan bahwa *topic modeling* mampu mengidentifikasi tema-tema diskusi yang muncul dari percakapan masyarakat di media sosial secara efektif.

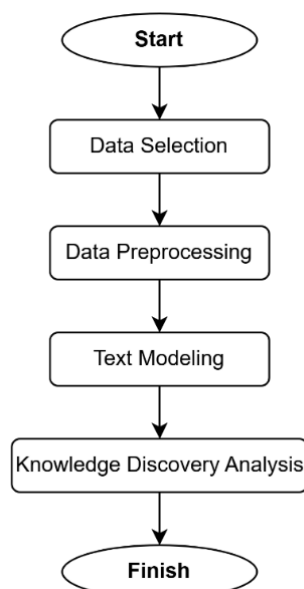
Penelitian lain yang menggunakan metode *Non-Negative Matrix Factorization* dilakukan oleh [14] yang menganalisis berita daring mengenai program makan bergizi gratis di Indonesia. Hasil penelitian tersebut menunjukkan bahwa metode NMF mampu mengidentifikasi beberapa topik utama seperti kebijakan pemerintah, distribusi program, serta dampak sosial dari program tersebut. Hasil tersebut menunjukkan bahwa metode NMF memiliki potensi yang baik dalam menghasilkan topik yang lebih terstruktur dan mudah diinterpretasikan.

Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa teknik *topic modeling* memiliki

peran penting dalam proses *knowledge discovery* pada data teks tidak terstruktur, khususnya pada data media sosial. Namun demikian, sebagian besar penelitian masih berfokus pada metode LDA, sedangkan penerapan metode *Non-Negative Matrix Factorization* (NMF) pada analisis informasi kesehatan di media sosial masih relatif terbatas. Oleh karena itu, penelitian ini memfokuskan pada penerapan metode NMF dalam proses *knowledge discovery* terhadap informasi kesehatan di Media X guna menghasilkan pemetaan topik yang lebih koheren, stabil, dan mudah diinterpretasikan.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksploratif dengan metode *text mining* untuk melakukan *knowledge discovery* terhadap informasi kesehatan di Media X (Twitter). Tahapan penelitian meliputi pengumpulan data, *preprocessing* data, penerapan *topic modeling* berbasis *Non-Negative Matrix Factorization* (NMF), serta evaluasi dan analisis hasil. Alur proses penelitian dapat terlihat pada Gambar 1.



Gambar 1. Alur Penelitian

3.1. Data Collection

Data diperoleh dari Media X menggunakan teknik web *crawling* dengan kata kunci terkait kesehatan. Rentang waktu pengambilan data adalah November 2025 hingga Februari 2026. Tahap awal menghasilkan 1.350 tweet dan setelah proses pembersihan diperoleh 1.338 tweet yang digunakan dalam analisis.

3.2. Data Preprocessing

Tahap *preprocessing* dilakukan untuk membersihkan data teks dan mengurangi *noise* agar hasil *topic modeling* menjadi lebih koheren dan mudah diinterpretasikan. Tahapan *preprocessing* yang dilakukan meliputi:

- Cleaning**
Menghapus elemen yang tidak relevan, seperti URL, hashtag, mention, emoji, tanda baca, angka, dan spasi berlebih.
- Case Folding**
Mengubah seluruh teks menjadi huruf kecil agar tidak terjadi perbedaan makna akibat variasi kapitalisasi.
- Normalization**
Menstandarkan kata-kata tidak baku atau bahasa informal ke bentuk baku, misalnya mengubah kata singkatan atau slang yang umum digunakan di media sosial.
- Tokenization**
Memecah teks menjadi unit kata (token) berdasarkan spasi atau pemisah tertentu.
- Stopword Removal**
Menghapus kata-kata umum yang tidak memiliki makna signifikan dalam penentuan topik, seperti kata hubung dan kata ganti. Selain itu, kata-kata yang sering muncul tetapi tidak relevan dengan konteks topik kesehatan juga dihapus.
- Stemming**
Mengubah kata berimbuhan menjadi bentuk kata dasar dengan menghapus awalan dan akhiran. Proses ini bertujuan untuk menyatukan variasi kata agar analisis topik menjadi lebih akurat dan konsisten.

Tahapan *preprocessing* ini bertujuan untuk menghasilkan data teks yang bersih dan siap digunakan dalam proses *topic modeling*. Hasil dari masing-masing tahapan dalam proses *preprocessing* data ditampilkan pada Tabel 2 untuk memperlihatkan transformasi data awal yang masih mentah hingga menjadi data yang telah terstruktur dan siap digunakan dalam tahap analisis selanjutnya.

3.3. Topic Modeling

Topic modeling dilakukan menggunakan metode *Non-Negative Matrix Factorization* (NMF). Pada tahap ini, data teks yang telah diproses terlebih dahulu direpresentasikan dalam bentuk *Term Frequency-Inverse Document Frequency* (TF-IDF). Representasi ini digunakan untuk membentuk matriks dokumen-kata.

Model *Non-Negative Matrix Factorization* (NMF) diimplementasikan menggunakan pustaka *scikit-learn* pada bahasa pemrograman Python. Penentuan jumlah topik dilakukan melalui proses eksperimen dengan beberapa variasi jumlah topik ($k = 3$ hingga $k = 10$). Setiap model kemudian dievaluasi menggunakan *topic coherence score* untuk menentukan konfigurasi jumlah topik yang paling optimal. Berdasarkan hasil evaluasi, nilai *coherence* tertinggi diperoleh pada $k = 5$ sehingga digunakan sebagai jumlah topik pada penelitian ini. Representasi teks dilakukan menggunakan pembobotan TF-IDF sebelum dimasukkan ke dalam model NMF untuk mengekstraksi struktur topik dari kumpulan dokumen.

NMF kemudian digunakan untuk memfaktorkan matriks tersebut menjadi dua matriks non-negatif, yaitu matriks dokumen topik dan matriks topik kata. Melalui matriks topik kata, kata-kata dengan bobot tertinggi pada setiap topik dapat diidentifikasi sebagai representasi topik tersebut. Jumlah topik diuji dalam rentang 5 untuk memperoleh konfigurasi yang optimal. Metode NMF dipilih karena mampu menghasilkan topik yang lebih koheren dan mudah diinterpretasikan, terutama pada data teks pendek dan informal seperti unggahan media sosial.

3.4. Knowledge Discovery Analysis

Tahap akhir penelitian adalah *knowledge discovery*, yaitu analisis hasil *topic modeling* untuk memperoleh wawasan dan pengetahuan baru. Analisis ini mencakup:

- Identifikasi topik informasi kesehatan yang paling dominan
- Pola kebutuhan informasi kesehatan Masyarakat
- Topik yang jarang dibahas namun memiliki tingkat kepentingan tinggi

- Tren pembahasan informasi kesehatan berdasarkan waktu

4. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah proses *Knowledge Discovery* terhadap informasi kesehatan yang diperoleh dari media sosial Twitter (X) menggunakan algoritma *Non-Negative Matrix Factorization* (NMF). Data yang digunakan berupa tweet berbahasa Indonesia yang mengandung kata kunci terkait isu kesehatan, yang kemudian diproses menggunakan bahasa pemrograman Python.

4.1. Data Collection

Data Collection dilakukan dengan cara web crawling dengan kata kunci terkait kesehatan pada aplikasi “X”. Pengumpulan data dilakukan dalam rentang waktu tertentu untuk memperoleh representasi diskusi kesehatan masyarakat yang aktual. Data yang berhasil dikumpulkan sebanyak 1.350 data dan diperoleh 1338 data setelah dilakukan pembersihan. Dataset awal yang diperoleh melalui teknik crawling ditampilkan pada Tabel 1 di bawah ini.

Tabel 1. Dataset awal hasil *web crawling*

No	Tanggal	Tweet
1	2026-02-10 07:50:45	Dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis
2	2025-11-28 04:12:06	@ntbd97__ Setuju banget Betul Rokok ilegal bukan cuma soal harga murah tapi risiko besar bagi kesehatan dan perlindungan konsumen. Edukasi dan pengawasan harus dilakukan #StopRokokIlegal

4.2. Data Preprocessing

Di tahap ini, data yang telah dikumpulkan diproses melalui *preprocessing* untuk membersihkan dan menormalkan teks agar optimal dalam pemodelan topik menggunakan algoritma *Non-Negative Matrix Factorization* (NMF). Proses *preprocessing* dilakukan menggunakan beberapa pustaka Python, antara lain Sastrawi untuk stemming bahasa Indonesia, NLTK untuk *tokenization* dan *stopword removal*, serta scikit-learn untuk pembobotan TF-IDF dan implementasi model NMF.

4.3. Cleaning

Data hasil *crawling* masih mengandung *noise*. Proses *cleansing* dilakukan untuk menghapus elemen tersebut sehingga teks menjadi lebih bersih dan siap digunakan pada proses selanjutnya. Hasil dari proses *cleansing* tersebut ditampilkan pada Tabel 2.

Tabel 2. *Cleaning*

Sebelum <i>Cleaning</i>	Sesudah <i>Cleaning</i>
Dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep	Dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep

Sebelum <i>Cleaning</i>	Sesudah <i>Cleaning</i>
ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis	ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis

4.4. Case Folding

Pada tahap ini dilakukan proses *case folding*, yaitu proses normalisasi teks dengan mengubah seluruh huruf menjadi huruf kecil (*lowercase*). Hasil dari proses *case folding* tersebut ditampilkan pada Tabel 3.

Tabel 3. *Case Folding*

Sebelum <i>Case Folding</i>	Sesudah <i>Case Folding</i>
Dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis	dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis

4.5. Normalization

Tahap *normalization* bertujuan untuk menstandarkan kata tidak baku atau slang menjadi bentuk kata baku sesuai dengan kaidah bahasa Indonesia, misalnya “udh” menjadi “sudah”, “kalo” menjadi “kalau”, dan “emng” menjadi “memang”. Hasil dari proses *normalization* tersebut ditampilkan pada Tabel 4.

Tabel 4. Normalization

Sebelum Normalization	Sesudah Normalization
dia cuma butuh di dengar dan dia pun sebenarnya udh tau solusi dari masalahnya apa untuk itu penanganan awal emng baiknya ke psikolog dulu karena nanti kalo pun tetep ke psikiater psikiater nya pasti liat riwayat nya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis	dia cuma butuh di dengar dan dia pun sebenarnya sudah tau solusi dari masalahnya apa untuk itu penanganan awal memang baiknya ke psikolog dulu karena nanti kalau pun tetap ke psikiater psikiater ya pasti lihat riwayat ya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis

4.6. Tokenization

Tahap *tokenization* bertujuan untuk memecah teks menjadi unit-unit kata atau token yang lebih kecil sehingga setiap kata dalam kalimat dapat dianalisis secara terpisah dalam proses pengolahan teks. Hasil dari proses *tokenization* tersebut ditampilkan pada Tabel 5.

Tabel 5. Tokenization

Sebelum Tokenization	Sesudah Tokenization
dia cuma butuh di dengar dan dia pun sebenarnya sudah tau solusi dari masalahnya apa untuk itu penanganan awal memang baiknya ke psikolog dulu karena nanti kalau pun tetap ke psikiater psikiater ya pasti lihat riwayat ya dulu dia kenapa dan meskipun nantinya ada perbedaan diagnosis	['dia', 'cuma', 'butuh', 'di', 'dengar', 'dan', 'dia', 'pun', 'sebenarnya', 'sudah', 'tau', 'solusi', 'dari', 'masalahnya', 'apa', 'untuk', 'itu', 'penanganan', 'awal', 'memang', 'baiknya', 'ke', 'psikolog', 'dulu', 'karena', 'nanti', 'kalau', 'pun', 'tetap', 'ke', 'psikiater', 'psikiater', 'ya', 'pasti', 'lihat', 'riwayat', 'ya', 'dulu', 'dia', 'kenapa', 'dan', 'meskipun', 'nantinya', 'ada', 'perbedaan', 'diagnosis']

4.7. Stopword Removal

Tahap *stopword removal* dilakukan untuk menghapus kata-kata umum yang sering muncul namun tidak memiliki makna *penting* dalam analisis teks. Dalam konteks ini, kata-kata yang tidak relevan terkait kesehatan juga dihapus. Hasil dari proses *stopword removal* tersebut ditampilkan pada Tabel 6.

Tabel 6. Stopword Removal

Sebelum Stopword Removal	Sesudah Stopword Removal
['dia', 'cuma', 'butuh', 'di', 'dengar', 'dan', 'dia', 'pun', 'sebenarnya', 'sudah', 'tau', 'solusi', 'dari', 'masalahnya', 'apa', 'untuk', 'itu', 'penanganan', 'awal', 'memang', 'baiknya', 'ke', 'psikolog', 'dulu', 'karena', 'nanti', 'kalau', 'pun', 'tetap', 'ke', 'psikiater', 'psikiater', 'ya', 'pasti', 'lihat', 'riwayat', 'ya', 'dulu', 'dia', 'kenapa', 'dan', 'meskipun', 'nantinya', 'ada', 'perbedaan', 'diagnosis']	['butuh', 'dengar', 'solusi', 'penanganan', 'baiknya', 'psikolog', 'psikiater', 'psikiater', 'lihat', 'riwayat', 'perbedaan', 'diagnosis']

4.8. Stemming

Pada tahap akhir ini, kata-kata berimbuhan diubah menjadi bentuk dasar untuk menyeragamkan variasi kata dalam proses analisis teks. Hasil dari proses *Stemming* tersebut ditampilkan pada Tabel 7.

Tabel 7. Stemming

Sebelum Stemming	Sesudah Stemming
['butuh', 'dengar', 'solusi', 'penanganan', 'baiknya', 'psikolog', 'psikiater', 'psikiater', 'lihat', 'riwayat', 'perbedaan', 'diagnosis']	butuh dengar solusi tangan baik psikolog psikiater psikiater lihat riwayat beda diagnosis

4.9. Topic Modeling

Topic modeling pada penelitian ini dilakukan menggunakan metode *Non-Negative Matrix Factorization* (NMF) untuk mengidentifikasi pola pembahasan informasi kesehatan yang dibagikan oleh pengguna media sosial X. Data teks yang telah melalui tahap *preprocessing* kemudian direpresentasikan menggunakan metode TF-IDF sehingga terbentuk matriks dokumen–kata sebagai input model NMF.

Berdasarkan hasil pemodelan, diperoleh lima topik utama yang merepresentasikan distribusi informasi kesehatan dalam dataset penelitian. Setiap topik dibentuk dari kumpulan kata dengan bobot tertinggi pada matriks topik–kata yang dihasilkan oleh model NMF. Karakteristik dari setiap topik tersebut beserta contoh tweet representatifnya disajikan pada Tabel 8 untuk memberikan gambaran konkret mengenai isi data. Kata-kata kunci yang muncul pada setiap kluster digunakan sebagai dasar interpretasi tema pembahasan pada masing-masing topik guna memahami kebutuhan informasi kesehatan pengguna secara lebih mendalam.

Tabel 8. Hasil interpretasi topik

No Topic	Label Topik	Tweet
0	Kesehatan Mental	Banyak remaja tidak menyadari bahwa tindakan mengejek memperlakukan atau mengintimidasi teman sebaya merupakan bentuk kekerasan psikologis yang dapat meninggalkan luka mendalam dalam jangka panjang.
1	Self-diagnosis dan konsultasi profesional	@tanyarlless Koar koar mental health tu buat mengenali emosi yang dirasakan diri sendiri belajar mengenal dan menghargai diri sendiri. Tapi wajib diingat gaboleh self diagnosis tetep wajib bimbingan profesional. Pikiran primitif kayak gini cant relate
2	Konsultasi Medis dan Layanan Kesehatan	@s___soulpeach kak mjb yaaa aku nemu di search bar aku konsul online dan di diagnosis bpd tp masih belum yakin jadi mau coba ke psikiater offline kakak dom jabodetabek bukannya??? ada saran gak ya psikiaternya anw semoga lekas membaik yaa kakk
3	Klinis & Mitos	Jujur kamu masih percaya mitos kesehatan yang mana? Mandi malam bikin sakit Kerokan menyembuhkan masuk angin Minum es bikin gemuk
4	Biaya dan akses pengobatan	Sedih banget temen sekelas anak ku Qoddarallah di diagnose kena kanker darah guys yang baca twit ini minta doa nya ya semoga temen anak ku di beri kesembuhan dan di mudahkan pengobatan nya

Hasil *topic modeling* menunjukkan bahwa diskursus kesehatan di media sosial tidak hanya berfokus pada informasi medis formal, tetapi juga mencerminkan pengalaman personal pengguna terkait kondisi kesehatan mereka. Hal ini menunjukkan bahwa media sosial dapat menjadi sumber data yang berharga untuk memahami kebutuhan informasi kesehatan masyarakat secara lebih luas.

Selain itu, hasil distribusi dokumen menunjukkan bahwa beberapa topik memiliki dominasi pembahasan yang lebih tinggi dibandingkan topik lainnya. Topik dominan merepresentasikan jenis isu atau informasi kesehatan yang paling sering dibahas oleh pengguna media sosial. Berdasarkan visualisasi distribusi topik, topik 0 memiliki proporsi terbesar sebesar 26,60%, diikuti oleh topik 2 sebesar 26,00%, topik 1 sebesar 21,67%, topik 4 sebesar 13,30%, dan topik 3 sebesar 12,40% dari keseluruhan data.

Temuan ini menunjukkan bahwa diskursus informasi kesehatan di media sosial tidak tersebar secara merata, melainkan cenderung terfokus pada beberapa isu kesehatan tertentu yang dianggap lebih relevan atau lebih sering dibicarakan oleh masyarakat.

4.10. Knowledge Discovery Analysis

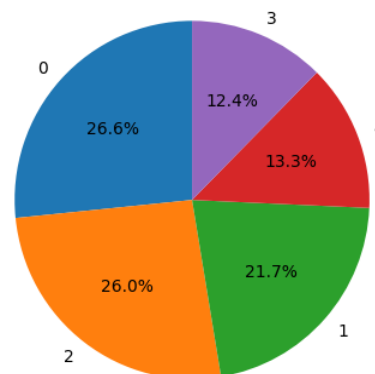
Tahap selanjutnya adalah *knowledge discovery*, yaitu proses interpretasi hasil *topic modeling* untuk memperoleh wawasan baru terkait pola diskursus dan penyebaran informasi kesehatan di media sosial.

4.11. Pola Topik Informasi Kesehatan Dominan

Analisis menunjukkan bahwa topik dengan persentase tertinggi menjadi indikator utama jenis informasi kesehatan yang paling sering dibagikan. Dominasi topik tertentu menunjukkan bahwa diskursus kesehatan di media sosial cenderung terfokus pada isu yang berkaitan dengan pengalaman personal dan kondisi psikologis pengguna. Hal ini mengindikasikan bahwa media sosial tidak hanya berfungsi sebagai sumber informasi kesehatan, tetapi juga sebagai ruang ekspresi dan dukungan sosial bagi

pengguna yang mengalami permasalahan kesehatan. Berdasarkan grafik pada Gambar 2, dapat dilihat bahwa distribusi persentase topik menunjukkan adanya satu atau dua topik yang memiliki proporsi paling dominan dibandingkan topik lainnya.

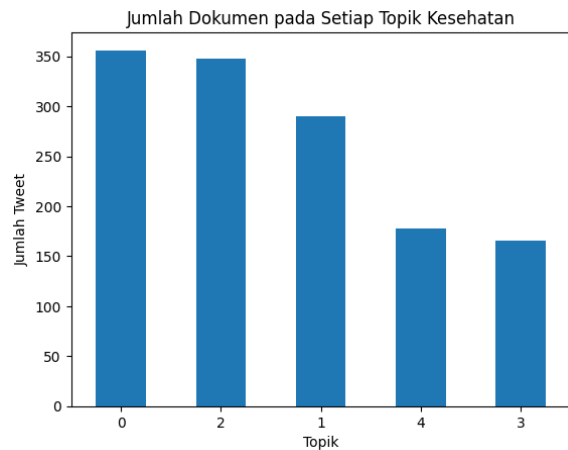
Distribusi Topik Informasi Kesehatan oleh Tenaga Medis



Gambar 2. Distribusi persentase topik informasi Kesehatan

4.12. Pola Kebutuhan Informasi Kesehatan Masyarakat

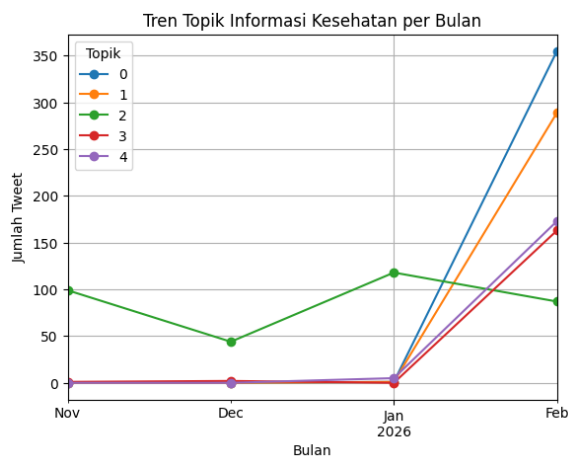
Distribusi topik memperlihatkan bahwa masyarakat cenderung membutuhkan informasi kesehatan yang bersifat praktis, mudah dipahami, dan dapat langsung diterapkan. Hal ini terlihat dari tingginya jumlah unggahan pada topik yang berkaitan dengan keluhan kesehatan umum dan edukasi preventif. Media sosial berperan sebagai sarana *knowledge sharing* yang mempertemukan kebutuhan informasi masyarakat dengan penyampaian edukasi oleh tenaga medis. Pada Gambar 3 dapat dilihat jumlah dokumen pada setiap topik kesehatan yang menunjukkan perbedaan frekuensi pembahasan antar topik.



Gambar 3. Jumlah dokumen pada setiap topik Kesehatan

4.13. Analisis Tren Topik Berdasarkan Waktu

Analisis temporal berdasarkan waktu unggahan menunjukkan adanya perubahan intensitas pembahasan topik pada setiap periode pengamatan. Peningkatan jumlah diskusi pada periode tertentu mengindikasikan adanya pengaruh faktor eksternal, seperti perubahan musim, kondisi lingkungan, maupun meningkatnya perhatian masyarakat terhadap isu kesehatan tertentu. Tren ini menunjukkan bahwa penyebaran informasi kesehatan di media sosial bersifat dinamis mengikuti kondisi aktual masyarakat. Dari grafik pada Gambar 4, dapat dilihat tren pembahasan topik berdasarkan waktu yang memperlihatkan adanya kenaikan dan penurunan jumlah unggahan pada periode tertentu

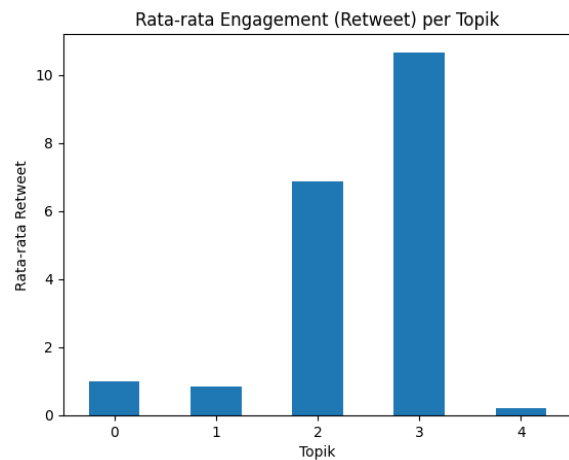


Gambar 4. Tren pembahasan topik informasi kesehatan berdasarkan waktu

4.14. Analisis Engagement Informasi Kesehatan

Selain distribusi topik, penelitian juga menganalisis tingkat interaksi pengguna melalui rata-rata jumlah retweet pada setiap topik. Hasil analisis menunjukkan bahwa beberapa topik memperoleh tingkat engagement yang lebih tinggi dibandingkan lainnya. Hal ini mengindikasikan bahwa jenis informasi kesehatan tertentu memiliki daya tarik lebih

besar serta lebih mudah disebarluaskan oleh pengguna media sosial. Berdasarkan Gambar 5, terlihat rata-rata engagement pada masing-masing topik kesehatan.



Gambar 5. Rata-rata engagement setiap topik

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa metode *Non-Negative Matrix Factorization* (NMF) efektif digunakan dalam proses *knowledge discovery* untuk mengidentifikasi pola kebutuhan informasi kesehatan masyarakat di media sosial X. Berdasarkan hasil pengujian, model NMF berhasil mengekstraksi lima topik utama yang mencakup kesehatan mental, *self-diagnosis*, edukasi medis, informasi klinis, serta biaya dan akses pengobatan. Analisis distribusi topik menunjukkan bahwa diskusi didominasi oleh isu kesehatan mental dan kondisi emosional, yang mengindikasikan bahwa media sosial X menjadi ruang bagi pengguna untuk mengekspresikan pengalaman dan keluhan terkait kesehatan psikologis. Selain itu, model NMF mampu menghasilkan topik yang koheren dan mudah diinterpretasikan pada data teks pendek seperti cuitan media sosial. Meskipun demikian, penelitian ini masih memiliki keterbatasan pada tahap preprocessing dan cakupan analisis. Oleh karena itu, penelitian selanjutnya disarankan untuk menerapkan teknik *preprocessing* yang lebih mendalam, seperti penanganan bahasa gaul atau slang yang lebih spesifik, serta mengintegrasikan analisis sentimen atau analisis emosi untuk memperoleh pemahaman yang lebih komprehensif terhadap diskursus kesehatan di media sosial.

DAFTAR PUSTAKA

- [1] S. Neely, C. Eldredge, and R. Sanders, "Health Information Seeking Behaviors on Social Media During the COVID-19 Pandemic Among American Social Networking Site Users : Survey Study," *J. Med. Internet Res.*, vol. 23, no. 6, pp. 1–10, 2021, doi: <https://www.jmir.org/2021/6/e29802>.
- [2] C. Jia and H. Qi, "Users' health information sharing behavior in social media: an integrated model," *Humanit. Soc. Sci. Commun.*, vol. 11,

- no. 1647, pp. 1–10, 2024, doi: 10.1057/s41599-024-04188-7.
- [3] F. A. Cahaya, W. Sari, D. A. Rambe, and Z. Lingga, “Peran Sosial Media dalam Promosi Kesehatan : Kajian Literatur,” *J. Riset, Inov. dan Teknol. Kabupaten Batang*, vol. 8, no. 2, pp. 60–67, 2024, doi: <https://doi.org/10.55686/ristek.v8i2.157>.
- [4] M. I. A. Rois, G. Dwilestari, and N. Suarna, “PREDIKSI PERSETUJUAN PINJAMAN MENGGUNAKAN DATASET LOAN APPROVAL MENGGUNAKAN ALGORITMA KLASIFIKASI,” *J. Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1342–1347, 2025, doi: <https://doi.org/10.36040/jati.v9i1.12486>.
- [5] M. Hutsebaut-buysse, K. Mets, and S. Latré, “Hierarchical Reinforcement Learning: A Survey and Open Research Challenges,” *Mach. Learn. Knowl. Extr.*, vol. 4, no. 1, pp. 172–221, 2022, doi: <https://doi.org/10.3390/make4010009>.
- [6] G. C. Kalangie and I. R. H. T. Tangkawarow, “Studi Literatur Penerapan Topic Modeling Menggunakan Latent Dirichlet Allocation (Lda) Dalam Analisis Dokumen Dan Laporan Publik,” *J. Pendidik. Teknol. Inf. dan Komun.*, vol. 6, no. 1, pp. 281–290, 2026.
- [7] S. Basuki and N. M. Fadillah, “Implementasi Algoritma Topic Modeling pada Abstrak Paper Ilmiah untuk Deteksi Tren Topik Penelitian,” *J. Repos.*, vol. 7, no. 1, pp. 7–22, 2025, doi: <https://doi.org/10.22219/repositor.v7i1.34464>.
- [8] M. R. Maulizidan and K. D. Tania, “Knowledge Discovery Based on Sentiment Analysis of Public Perceptions About Generative AI on X,” *Indones. J. Informatics Educ.*, vol. 9, no. 2, 2025, doi: <http://dx.doi.org/10.20961/ijie.v9i2.107758>.
- [9] A. Muhaimin *et al.*, “Social Media Analysis and Topic Modeling : Case Study of Stunting in Indonesia,” *J. Inform. dan Teknol. Inf.*, vol. 20, no. 3, pp. 406–415, 2023, doi: 10.31515/telematika.v20i3.10797.
- [10] M. Jelita, “Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI,” *Media Pustak.*, vol. 31, no. 3, pp. 253–265, 2024, doi: 10.37014/medpus.v31i3.5237.
- [11] D. Ridhwanulah and D. H. Fudholi, “Pemodelan Topik pada Cuitan tentang Penyakit Tropis di Indonesia dengan Metode Latent Dirichlet Allocation,” *J. Ilm. Sinus*, vol. 20, no. 1, pp. 1–22, 2022, doi: <http://dx.doi.org/10.30646/sinus.v20i1.589>.
- [12] B. D. Santosa, N. Fatimah, N. I. Kusumaningtyas, U. S. Aesyti, and H. Santoso, “Indonesian Journal on Data Science Analisis Kepercayaan Masyarakat Tentang Kepolisian Indonesia di Twitter Menggunakan Latent Dirichlet Allocation (LDA),” *Indoensian J. Data Sci.*, vol. 1, no. 2, pp. 66–76, 2023, doi: <https://doi.org/10.30989/ijds.v1i2.1198>.
- [13] D. K. Harahap, K. D. Tania, and P. E. Sevtiyuni, “Topic Mining-Based Knowledge Discovery of User Health Information Needs,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 4, pp. 641–653, 2025, doi: <https://doi.org/10.35882/ijeemi.v7i4.270>.
- [14] I. Dwijayanti, A. R. Lahitani, and M. Habibi, “Topic Analysis of Indonesian Online News on the Free Nutritious Meal Program Using Non-Negative Matrix Factorization,” *Compiler*, vol. 14, no. 2, pp. 119–130, 2025, doi: 10.28989/compiler.v14i2.3499.