

Pendekatan Hibrid Berbasis NLP untuk Deteksi Emosi dan Kesehatan Mental di X.com dengan TF-IDF dan BERTopic yang Dioptimasi

Abdullah Ahmad^{1*}, M. Safii², Sundari Retno Andani³, Anjar Wanto⁴

¹Informatika, STIKOM Tunas Bangsa, Kota Pematangsiantar, Sumatera Utara, Indonesia

^{2,3,4}Informatika, STIKOM Tunas Bangsa, Kota Pematangsiantar, Sumatera Utara, Indonesia

e-mail: ¹abdul@amiktunasbangsa.ac.id, ²m.safii@amiktunasbangsa.ac.id,

³sundari.ra@amiktunasbangsa.ac.id, ⁴anjarwanto@amiktunasbangsa.ac.id

Abstrak

Penelitian ini dilakukan untuk menganalisis dan membandingkan efektivitas dua pendekatan pemodelan topik berbasis pemrosesan bahasa alami dalam mendeteksi emosi dan kesehatan mental pada teks dari platform *X.com*. Model dasar dibangun menggunakan representasi *Term Frequency-Inverse Document Frequency* yang diintegrasikan dengan algoritma *BERTopic*. Selanjutnya, dilakukan optimasi parameter pada komponen reduksi dimensi dan klusterisasi untuk meningkatkan koherensi topik. Hasil pengujian menunjukkan bahwa model yang dioptimasi menghasilkan peningkatan nilai koherensi dari 0,327 menjadi 0,398, yang menandakan hubungan semantik antar kata dalam satu topik menjadi lebih kuat dan konsisten. Berdasarkan analisis, pendekatan berbasis konteks terbukti lebih unggul dalam menghasilkan topik yang representatif dan relevan terhadap ekspresi emosi serta isu kesehatan mental di media sosial.

Kata kunci: Pemrosesan Bahasa Alami, Pemodelan Topik, *BERTopic*, *TF-IDF*, Kesehatan Mental

Abstract

This study was conducted to analyze and compare the effectiveness of two natural language processing-based topic modeling approaches in detecting emotions and mental health in text from the X.com platform. The baseline model was built using the Term Frequency-Inverse Document Frequency representation integrated with the BERTopic algorithm. Next, parameter optimization was performed on the dimensionality reduction and clustering components to improve topic coherence. Test results showed that the optimized model resulted in an increase in coherence value from 0.327 to 0.398, indicating stronger and more consistent semantic relationships between words within a topic. Based on the analysis, the context-based approach proved superior in generating representative and relevant topics related to emotional expressions and mental health issues on social media.

Keywords: Natural Language Processing, Topic Modeling, *BERTopic*, *TF-IDF*, Mental Health

1. Pendahuluan

Perkembangan media sosial telah membawa perubahan besar dalam cara manusia berkomunikasi dan mengekspresikan emosi [1], [2], [3], [4]. Platform seperti *X.com* (sebelumnya *Twitter*) menjadi ruang bagi individu untuk menyampaikan perasaan [5], [6], [7], opini [8], [9], [10], dan pengalaman pribadi secara spontan dan terbuka [11], [12], [13]. Unggahan-unggahan tersebut mencerminkan dinamika psikologis masyarakat modern, termasuk kondisi emosional dan kesehatan mental mereka [14], [15], [16], [17]. Namun, besarnya volume data yang dihasilkan serta keragaman gaya bahasa informal membuat proses analisis manual menjadi tidak efisien [18], [19], [20]. Oleh karena itu, pendekatan berbasis *Natural Language*

Processing (NLP) menjadi solusi yang relevan untuk memahami ekspresi emosi di media sosial secara otomatis dan komprehensif [21], [22], [23], [24], [25].

Berbagai penelitian sebelumnya telah memanfaatkan teknik *NLP* dalam mendeteksi emosi dan menganalisis kesehatan mental di media sosial. Beberapa penelitian terdahulu yang dimaksud seperti yang dilakukan oleh Cahyani dan Patasik (2021) yang membandingkan metode *TF-IDF* dan *Word2Vec* dalam klasifikasi emosi pada data Twitter, dan hasilnya menunjukkan bahwa *TF-IDF* memiliki akurasi dan *F1-score* lebih tinggi, terutama pada dataset kecil. Temuan ini menunjukkan bahwa pendekatan berbasis frekuensi kata masih efektif, meskipun terbatas dalam memahami konteks semantic [26]. Selanjutnya, Monica Yoshe Dhinora dan Evangs Mailoa (2025) meneliti ekspresi depresi mahasiswa melalui *tweet* menggunakan *TF-IDF* dan *Random Forest*, dengan akurasi 87,51% dan sekitar 68,1% *tweet* menunjukkan emosi negatif yang mengindikasikan stres atau depresi. Hasil tersebut memperkuat efektivitas *TF-IDF* dalam mengekstraksi fitur linguistik teks pendek, namun metode ini belum mampu menangkap makna kontekstual yang lebih mendalam [27]. Pendekatan konvensional seperti *Naïve Bayes*, *Support Vector Machine (SVM)*, dan representasi teks berbasis *TF-IDF* banyak digunakan karena kesederhanaan dan efisiensinya [28], [29], [30], [31]. Di sisi lain, pendekatan modern berbasis representasi kontekstual seperti *BERT* dan turunannya menunjukkan peningkatan performa dalam menangkap makna semantik yang lebih dalam. Meskipun demikian, sebagian besar penelitian terdahulu hanya menerapkan salah satu pendekatan secara terpisah, baik klasik maupun modern. Hal ini menimbulkan keterbatasan dalam menggabungkan kekuatan analisis semantik dari model modern dengan efisiensi komputasi dari metode tradisional.

Dari tinjauan literatur tersebut, dapat diidentifikasi adanya kesenjangan penelitian (*research gap*), yaitu belum adanya pendekatan hibrid yang mengombinasikan metode *TF-IDF* dan *BERTopic* untuk mendeteksi emosi dan kesehatan mental secara bersamaan dalam satu kerangka analisis yang terintegrasi dan dioptimasi. Penelitian sebelumnya cenderung berfokus pada klasifikasi emosi secara terpisah dari analisis kesehatan mental, padahal kedua aspek ini saling berkaitan dan dapat saling memperkuat dalam proses interpretasi hasil. Selain itu, belum banyak studi yang menerapkan model tersebut pada konteks data berbahasa Indonesia dari *X.com*, yang memiliki karakteristik linguistik khas seperti penggunaan slang, singkatan, dan campuran bahasa.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan pendekatan hibrid berbasis *NLP* yang menggabungkan kekuatan representasi statistik dari *TF-IDF* dengan kemampuan ekstraksi topik kontekstual dari *BERTopic* yang dioptimasi [26], [32], [33], [34]. Pendekatan ini diharapkan mampu meningkatkan akurasi dan relevansi dalam deteksi emosi serta memberikan pemetaan topik yang lebih bermakna terhadap kondisi kesehatan mental pengguna *X.com* [35], [36], [37], [38]. Penelitian ini bertujuan untuk mengembangkan model analisis yang mampu mendeteksi emosi sekaligus memetakan kecenderungan topik yang berkaitan dengan kesehatan mental berdasarkan data unggahan pengguna di platform tersebut.

Kebaruan (*novelty*) dari penelitian ini terletak pada penerapan kombinasi *TF-IDF* dan *BERTopic* dalam satu sistem analisis yang terintegrasi dan dioptimasi secara bersamaan. Pendekatan ini belum banyak dieksplorasi dalam penelitian sebelumnya, khususnya pada konteks bahasa Indonesia dan data *X.com*. Selain menghasilkan model yang lebih akurat, penelitian ini juga memberikan kontribusi dalam aspek interpretabilitas hasil, karena *BERTopic* memungkinkan analisis berbasis topik yang dapat dipahami secara konseptual oleh peneliti sosial maupun psikolog.

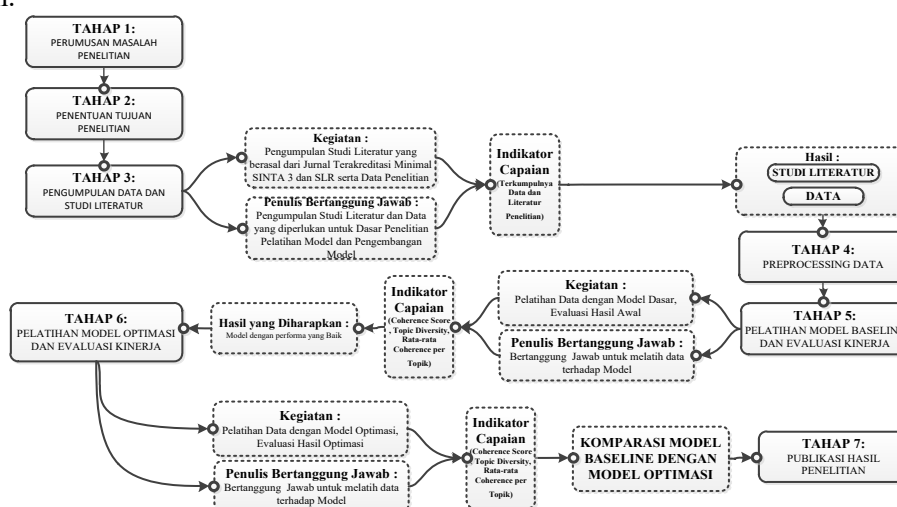
Penelitian ini memiliki batasan (*scope and limitations*) pada data publik *X.com* yang berbahasa Indonesia, dengan fokus pada teks unggahan yang mencerminkan ekspresi emosi. Elemen non-teks seperti gambar, emoji, dan metadata pengguna tidak disertakan dalam analisis. Selain itu, hasil penelitian ini tidak dimaksudkan untuk diagnosis klinis, melainkan sebagai indikasi awal terhadap kondisi emosional dan pola linguistik yang dapat berkaitan dengan kesehatan mental.

Secara keseluruhan, penelitian ini diharapkan memberikan kontribusi akademik dan praktis. Dari sisi akademik, penelitian ini memperluas kajian *NLP* dalam konteks analisis emosi dan kesehatan mental melalui pendekatan hibrid yang memadukan metode tradisional dan modern. Dari sisi praktis, hasil penelitian dapat dimanfaatkan untuk mendukung upaya deteksi dini potensi gangguan emosional di media sosial serta memberikan wawasan bagi lembaga kesehatan dan peneliti dalam memahami kondisi psikologis masyarakat digital secara lebih komprehensif.

2. Metode Penelitian

2.1. Tahapan Penelitian

Untuk memastikan hasil penelitian dapat dipertanggungjawabkan secara ilmiah, penelitian ini dilakukan secara empiris, logis, dan sistematis, mulai dari perumusan masalah hingga penyusunan laporan [39]. Gambar 1 menunjukkan struktur tahapan penelitian yang diusulkan.



Gambar 1. Tahapan Penelitian

2.1.1. Identifikasi Masalah (*Problem Identification*)

Tahap ini mencakup proses identifikasi dan perumusan masalah yang menjadi dasar pelaksanaan penelitian. Permasalahan utama yang diangkat adalah bagaimana meningkatkan kualitas analisis topik, emosi, dan kesehatan mental berbasis teks dari *platform X.com* menggunakan pendekatan *NLP*. Metode konvensional seperti *TF-IDF* sering menghasilkan topik yang kurang representatif karena hanya memperhatikan frekuensi kata tanpa memahami konteks semantik. Oleh karena itu, diperlukan pendekatan yang lebih adaptif dan kontekstual seperti *BERTopic*, yang menggabungkan *embedding* dari model *BERT* dengan algoritma reduksi dimensi dan klusterisasi untuk menghasilkan topik yang lebih bermakna.

2.1.2. Populasi, Sampel, dan Teknik Pengumpulan Data

Populasi dalam penelitian ini adalah seluruh unggahan pengguna pada *platform X.com* (sebelumnya *Twitter*) yang mengandung ekspresi emosi dan isu kesehatan mental. Sampel penelitian diambil sebanyak 518 data teks (*tweet*) berbahasa Indonesia. Teknik pengambilan sampel yang digunakan adalah purposive sampling, yaitu pemilihan data berdasarkan kata kunci tertentu yang relevan dengan konteks emosi dan kesehatan mental.

Pengumpulan data dilakukan melalui proses web scraping menggunakan *SNScrape*, dengan menerapkan kata kunci seperti “sedih”, “stres”, “cemas”, “depresi”, “lelah”, “capek”, “sendiri”, “*overthinking*”, “mental”, dan istilah lain yang berhubungan dengan gejala emosional. Data yang dikumpulkan merupakan unggahan publik tanpa interaksi dengan pengguna, dan mencakup periode September hingga Oktober 2025.

Data yang diperoleh terdiri dari beberapa atribut, seperti identitas tweet, teks unggahan (*full_text*), tanggal unggahan (*created_at*), dan atribut tambahan seperti jumlah *retweet* dan favorit. Seluruh data kemudian diseleksi dan difilter untuk memastikan kesesuaian dengan kriteria penelitian sebelum masuk ke tahap *preprocessing*.

2.1.3. Tujuan Penelitian (*Research Objectives*)

Tujuan penelitian ini adalah mengembangkan dan mengevaluasi pendekatan hibrida berbasis *NLP* yang mengombinasikan *TF-IDF* dan *BERTopic* untuk mendeteksi emosi dan kesehatan mental pada data teks dari *X.com*. Selain itu, penelitian ini bertujuan untuk membandingkan performa model *baseline* (*TF-IDF*) dengan model optimasi (*BERTopic*) berdasarkan metrik koherensi dan interpretabilitas topik, sehingga dapat diperoleh model yang lebih efektif dalam analisis teks emosional dan psikologis.

2.1.4. Pengumpulan Data dan Studi Literatur (*Data and Literature Collection*)

Tahap pertama penelitian berfokus pada pengumpulan data dan studi literatur. Pada tahap ini, dilakukan pengumpulan data teks dari *platform X.com* yang relevan dengan konteks emosi dan kesehatan mental dengan teknik *Crawling*. Selain itu, peneliti melakukan kajian literatur terhadap penelitian-penelitian terdahulu dari jurnal terindeks SINTA atau *Scopus* sebagai dasar konseptual dan metodologis. Hasil tahap ini adalah dataset yang siap digunakan serta pemahaman teoritis terhadap metode *TF-IDF*, *BERTopic*, dan algoritma pendukung seperti *UMAP* dan *HDBSCAN*.

Berikut merupakan contoh data yang digunakan dalam penelitian :

Tabel 1. Dataset Penelitian

conversation_id_str	created_at	... full text
1,97305E+18	Tue Sep 30 15:56:33 +0000 2025	September tahun ini ternyata Allah buat aku jadi lebih kuat dan lebih kuat lagi. Sabar dan lebih sabar lagi MasyaAllah.. Tidak mudah melalui nya tapi aku percaya dibalik semua itu pasti ada hikmah dan keberkahan Aamiin. https://t.co/EZOe3ueG6C
1,97305E+18	Tue Sep 30 15:55:04 +0000 2025	Hai apa kabar imanmu? Sudahkah mengaji hari ini? Sudahkah mentadaburi quran? Sudahkah mengamalkannya? Karena kebahagiaan itu dipergilirkan. Maka hati yg kuat diperlukan saat kesedihan yg dipergilirkan. Tidak ada yg abadi di dunia yg fana smw sementara sj.
1,97305E+18	Tue Sep 30 15:40:57 +0000 2025	jikalau memang kuat ya akan ada jalanya. jikalau tidak ada jalanya berarti memang sudah takdir.
1,97305E+18	Tue Sep 30 15:39:52 +0000 2025	Pertempuran Kita Jika kita bertempur hanya dengan mengandalkan jumlah dan kekuatan maka mereka (musuh) lebih banyak jumlahnya dan lebih kuat dari kita. Dan pasti kita tidak akan mampu melawan mereka. Namun jika kita bertempur dengan mengandalkan kekuatan dari Allah https://t.co/OPpB0Bgnvg
1,97305E+18	Tue Sep 30 15:38:54 +0000 2025	@paaries26 Tidak mood km panasnya ga kuat banget hiks
1,97304E+18	Tue Sep 30 15:35:00 +0000 2025	@muncorner Kebetulan sampai sekarang masih begini. Aku nggak bisa ngasih solusinya cuma semisal emang udah enggak kuat disimpan sendiri coba diceritain ke orang yang kamu percaya. Urusan belakang direspon sesuai yang kamu harapkan atau tidak setidaknya kamu pernah mencoba
1,97304E+18	Tue Sep 30 15:17:34 +0000 2025	kehidupan memang tidak adil justru aneh jikalau selalu berjalan sesuai keinginan kita. banyak yang menginginkan kita gagal banyak juga yang mendoakan kita agar dapat mencapai impian. yang bisa kita lakukan hanya berusaha dan terus berdoa agar kita senantiasa kuat dan terbiasa.
1,97304E+18	Tue Sep 30 15:09:45 +0000 2025	@ganjarpranowo Kalo aku punya anak pondasi agama yg aku tekan kan Km kalo ujian datang di hidup mereka mereka tidak goyah dan kuat iman Islam. Karena faktanya hidup apa yg kita rencanakan belum tentu sesuai harapan. Takdir dan Lika likunya. Dunia.
1,97304E+18	Tue Sep 30 15:07:15 2025	sepertinya aku pensiun minum kopi lambung tidak kuat
...	...	...
1,97154E+18	Fri Sep 26 11:21:47 +0000 2025	Finora AI Expectation #ETHUSDT 1h - Untuk saat ini ETHUSDT cenderung masih BERISIK alias kemungkinan LEBIH MUDAH TURUN ke area 3830 3815 sebelum terjadi pantulan signifikan. Jika harga mampu break bawah 3815 dan tidak muncul reversal kuat target selanjutnya di 3739 17. - https://t.co/DZVfa6nmkx

2.1.5. Pre-Processing Data

Tahap ini dilakukan proses menghapus *URL*, mention, tagar, dan karakter *non-alfabet*, Menghapus emoji dengan pustaka emoji, *Case folding* yakni mengubah seluruh teks menjadi huruf kecil, *Stopword removal* dengan *stopwords* Bahasa Indonesia (*nltk*) dan *Stemming* menggunakan *Sastrawi*.

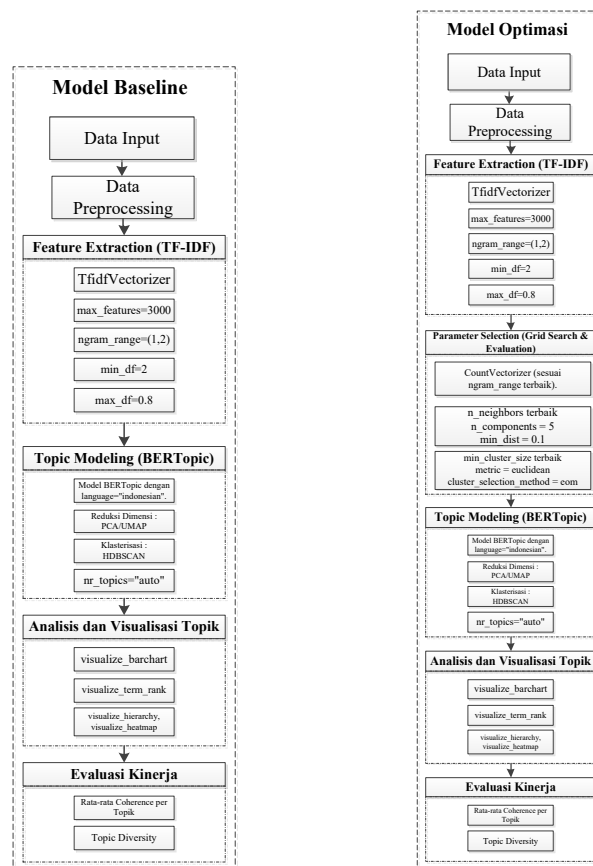
2.1.6. Pelatihan Model *Baseline* dan Evaluasi Awal (*Baseline Model Training and Evaluation*)

Tahap kedua meliputi proses pelatihan model baseline menggunakan pendekatan *TF-IDF* yang dikombinasikan dengan algoritma *UMAP* untuk reduksi dimensi dan *HDBSCAN* untuk pembentukan kluster topik. Model ini menghasilkan representasi awal topik yang akan digunakan sebagai dasar pembandingan dengan model optimasi. Evaluasi dilakukan menggunakan metrik *coherence* untuk menilai keterkaitan semantik antar kata dalam satu topik.

2.1.7. Pelatihan Model Optimasi dan Evaluasi Kinerja (*Optimized Model Training and Performance Evaluation*)

Tahap ketiga mencakup pengembangan model optimasi menggunakan *BERTopic*. Pada tahap ini, dilakukan penyesuaian parameter utama seperti jumlah tetangga (*n_neighbors*), komponen reduksi (*n_components*), dan ukuran kluster minimum (*min_cluster_size*) pada algoritma *UMAP* dan *HDBSCAN*. Tujuannya adalah untuk memperoleh konfigurasi model yang optimal dalam membentuk topik yang lebih konsisten dan mudah diinterpretasikan. Evaluasi kinerja dilakukan secara kuantitatif melalui metrik *coherence* dan secara kualitatif melalui analisis kesesuaian topik dengan konteks data.

Berikut merupakan perbandingan alur dan parameter yang digunakan untuk pengolahan data dengan *Baseline Model (TF-IDF + BERTopic)* serta *Model Optimasi (TF-IDF + BERTopic + Grid Search)*:



Gambar 2. Alur Model Optimasi

2.1.8. Komparasi Model *Baseline* dan Model Optimasi (*Comparison between Baseline and Optimized Model*)

Tahap keempat merupakan proses perbandingan menyeluruh antara model *baseline* dan model optimasi. Perbandingan dilakukan berdasarkan metrik evaluasi yang sama untuk memastikan objektivitas hasil. Analisis ini bertujuan menilai efektivitas penerapan optimasi pada *BERTopic* dibandingkan model *TF-IDF* dalam menghasilkan topik yang relevan dan representatif terhadap data emosi serta kesehatan mental pengguna *X.com*.

2.1.9. Publikasi dan Diseminasi Hasil Penelitian (*Publication and Research Dissemination*)

Tahap kelima adalah penyusunan laporan dan publikasi hasil penelitian. Seluruh proses mulai dari pengumpulan data, pelatihan model, hingga analisis komparatif didokumentasikan secara sistematis. Hasil penelitian kemudian disiapkan untuk dipublikasikan dalam jurnal atau prosiding ilmiah yang relevan dengan bidang *Natural Language Processing* dan *text mining*, agar dapat memberikan kontribusi akademik terhadap pengembangan metode analisis emosi dan kesehatan mental berbasis *NLP*.

2.1.10 Etika Penelitian & Privasi Data

Penelitian ini menggunakan data publik *X.com* dan mengikuti etika penelitian media sosial, yaitu tidak mengumpulkan data personal sensitif, tidak melakukan interaksi langsung dengan pengguna, dan hanya menggunakan teks yang bersifat publik serta dianonimkan untuk menjaga privasi.

3. Hasil dan Pembahasan

3.1. Hasil *Pre-Processing Data*

Pada tahap ini, dilakukan proses persiapan data sebelum digunakan dalam pemodelan topik. Dataset yang digunakan terdiri atas 518 data teks hasil pengumpulan dari *platform X.com* yang berkaitan dengan emosi dan kesehatan mental. Tahapan *preprocessing* meliputi *case folding*, pembersihan teks dari tautan, tagar, dan emoji, penghapusan kata tidak penting (*stopword removal*), serta *stemming* menggunakan pustaka *Sastrawi* agar setiap kata dikembalikan ke bentuk dasarnya.

Proses ini menghasilkan data teks yang lebih bersih dan seragam sehingga siap digunakan untuk tahap ekstraksi fitur dan pemodelan topik. Hasil akhir *preprocessing* ditunjukkan pada Tabel 2.

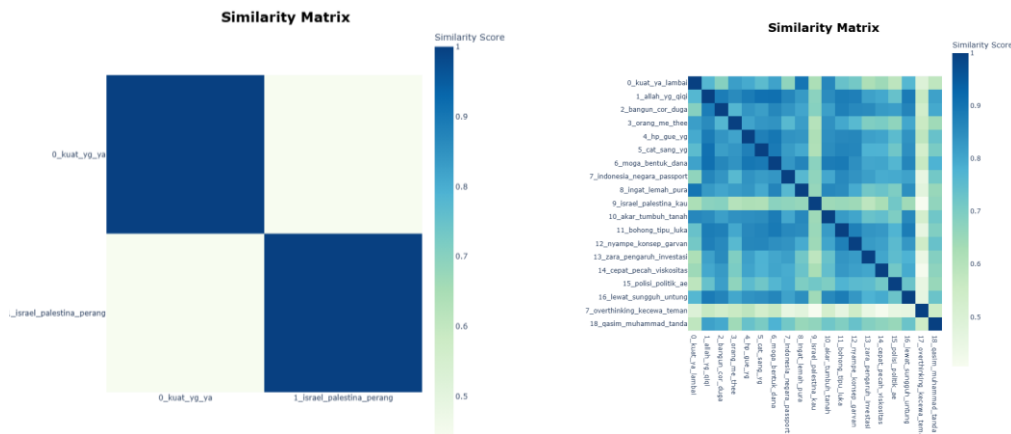
Tabel 2. Contoh Hasil *Preprocessing* Teks

No	Teks Asli	Hasil <i>Preprocessing</i>
1	September tahun ini ternyata Allah buat aku ja...	september allah kuat kuat sabar sabar masyaall...
2	Hai apa kabar imanmu? Sudahkah mengaji hari in...	hai kabar iman aji tadabur quran amal bahagia ...
3	jikalau memang kuat ya akan ada jalanya. jikal...	kuat ya jalani jalan takdir
4	Pertempuran Kita Jika kita bertempur hanya den...	pejuang andal kuat musuh kuat lawan temp...
5	@paaries26 Tidak mood krn panasnya ga kuat ban...	mood krn panas ga kuat banget hiks

Tahap ini bertujuan untuk mengurangi variasi linguistik dan memastikan bahwa model dapat mengenali pola semantik dengan lebih baik. Dengan data yang telah dinormalisasi ini, proses representasi dan pembentukan topik dapat dilakukan secara optimal.

3.2. Hasil *Similarity Topic*

Berikut merupakan hasil *topic word* untuk model *baseline* (a) dan model optimasi (b).

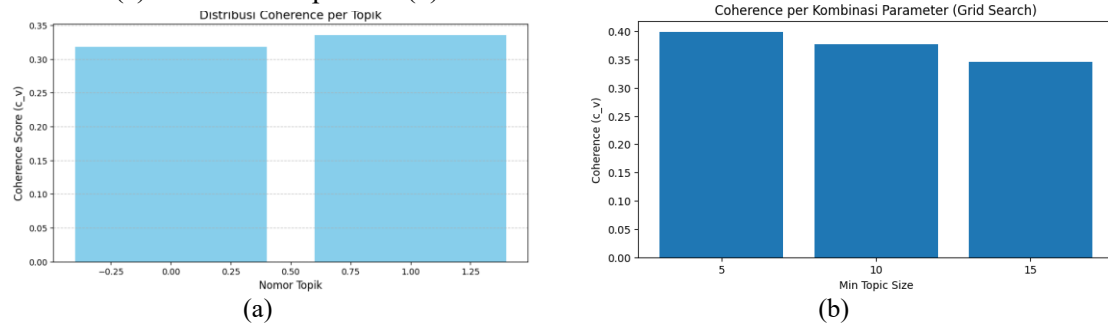


Gambar 3. *Similarity Topic Model Baseline (a) Model Optimasi (b)*

Gambar menunjukkan perbandingan matriks kesamaan antar topik pada model *baseline* dan model optimasi. Pada model *baseline* (a), terlihat bahwa hanya terdapat dua topik dengan pola hubungan yang sederhana dan jarak semantik yang relatif berdekatan, menandakan bahwa model belum mampu membedakan struktur makna antar topik secara jelas. Sebaliknya, model optimasi (b) menampilkan distribusi kesamaan yang lebih kompleks dengan variasi tingkat kemiripan di antara topik-topik yang terbentuk. Pola warna yang lebih bervariasi menunjukkan bahwa hasil optimasi berhasil menghasilkan topik yang lebih beragam namun tetap memiliki keterkaitan semantik yang terukur. Hal ini mengindikasikan peningkatan kemampuan model dalam memetakan hubungan makna antar tema, sekaligus memperkuat struktur topik yang terbentuk setelah optimasi dilakukan.

3.3. Coherence Score dan Topic Diversity

Berikut gambaran *Coherence Score* dan *Topic Diversity* yang dihasilkan untuk model *baseline* (a) serta model optimasi (b).



Gambar 4. *Coherence Score Model Baseline*

Hasil pemodelan menunjukkan bahwa model *baseline* menghasilkan *Coherence Score* sebesar 0.327, *Topic Diversity* sebesar 0.95, dan Rata-rata *Coherence* per Topik sebesar 0.327. Nilai tersebut menunjukkan bahwa hubungan semantik antar kata dalam satu topik masih tergolong moderat, meskipun variasi antar topik cukup tinggi. Sedangkan hasil pengujian menunjukkan peningkatan performa dengan *Coherence Score* sebesar 0.398, meningkat 0.071 dari *model baseline*. Hal ini menunjukkan bahwa hubungan semantik antar kata dalam topik menjadi lebih kuat dan konsisten.

3.4. Evaluasi dan Komparatif

Perbandingan antara model *baseline* dan model optimasi dilakukan berdasarkan metrik *coherence* dan kualitas interpretasi topik. Tabel 3 menunjukkan hasil komparasi kuantitatif antara kedua model.

Tabel 3. Perbandingan Model *Baseline* dan Model Optimasi

Aspek Evaluasi	Model Baseline	Model Optimasi
Representasi Teks	TF-IDF	CountVectorizer (optimasi)
Reduksi Dimensi	PCA	UMAP
Klasterisasi	HDBSCAN (default)	HDBSCAN (min_cluster_size optimal)

Hasil komparatif ini menunjukkan bahwa proses optimasi berhasil meningkatkan koherensi dan interpretabilitas topik secara signifikan. Model optimasi menghasilkan topik yang lebih mudah dipahami dan relevan dengan konteks emosi serta kesehatan mental pengguna media sosial.

3.5. Pembahasan

Hasil penelitian menunjukkan bahwa model berbasis konteks seperti *BERTopic* yang dioptimasi mampu menghasilkan topik yang lebih representatif dibandingkan pendekatan berbasis frekuensi seperti *TF-IDF*. Peningkatan nilai coherence sebesar 0.071 mengindikasikan bahwa optimasi parameter pada *UMAP* dan *HDBSCAN* berhasil memperkuat hubungan semantik antar kata dalam topik. Peningkatan ini terjadi karena *BERTopic* memanfaatkan representasi berbasis transformer embeddings yang mampu menangkap konteks semantik antar kata, sehingga menghasilkan kluster topik yang lebih konsisten. Proses optimasi pada parameter *UMAP* dan *HDBSCAN* juga meningkatkan pemisahan antar cluster dan kepadatan topik, sehingga distribusi kata menjadi lebih terstruktur dan mudah diinterpretasikan. Hal ini menunjukkan bahwa pendekatan berbasis embedding lebih efektif dalam memodelkan makna psikologis dalam teks pendek.

Temuan ini sejalan dengan penelitian Monica Yoshe Dhinora dan Evangs Mailoa (2025) yang menyatakan bahwa *BERTopic* memiliki kemampuan menghasilkan topik yang lebih stabil dan relevan dibandingkan pendekatan tradisional seperti *TF-IDF* dan *LDA* [27]. Namun, hasil ini sedikit berbeda dengan Cahyani dan Patasik (2021) yang menyimpulkan bahwa *TF-IDF* masih unggul pada tekstual pendek, sehingga dapat dikatakan bahwa keberhasilan *BERTopic* dalam penelitian ini dipengaruhi oleh proses optimasi parameter yang dilakukan [26].

Secara praktis, model optimasi menunjukkan potensi besar untuk diterapkan dalam analisis emosi dan kesehatan mental di media sosial, terutama untuk mendukung deteksi awal pola linguistik yang mengarah pada gangguan emosional. Meski demikian, penelitian ini memiliki keterbatasan pada jumlah sampel yang masih terbatas dan tidak melibatkan validasi ahli dari bidang psikologi klinis, sehingga hasil yang diperoleh belum dapat dijadikan indikator diagnostik klinis.

4. Kesimpulan

Penelitian ini bertujuan untuk mengembangkan pendekatan hibrid berbasis *NLP* dalam mendeteksi emosi dan kesehatan mental pada unggahan *X.com* melalui integrasi *TF-IDF* dan *BERTopic* yang dioptimasi. Berdasarkan hasil pengujian, pendekatan yang diusulkan berhasil meningkatkan kualitas pemodelan topik, ditunjukkan dengan peningkatan nilai coherence dari 0,327 pada model *baseline* menjadi 0,398 pada model optimasi. Hasil ini menunjukkan bahwa tujuan penelitian telah tercapai, yaitu memperoleh model yang lebih efektif dalam menghasilkan topik yang relevan dan representatif terhadap ekspresi emosi dan isu kesehatan mental.

Secara praktis, model yang dikembangkan memberikan kontribusi dalam menyediakan kerangka analisis yang lebih komprehensif untuk pemetaan emosi dan indikator kesehatan mental di media sosial. Pendekatan ini dapat dimanfaatkan oleh peneliti, praktisi kesehatan, maupun *platform* digital untuk mendukung deteksi dini potensi gangguan emosional berbasis teks. Dari sisi akademik, penelitian ini berkontribusi dalam pengembangan metode *NLP* dengan memadukan keunggulan representasi statistik dan kontekstual untuk meningkatkan kualitas topic modeling pada data teks informal berbahasa Indonesia.

Penelitian ini masih memiliki keterbatasan, seperti jumlah sampel yang relatif terbatas dan tidak melibatkan validasi ahli psikologi klinis dalam menilai interpretasi topik. Untuk penelitian selanjutnya, disarankan memperluas cakupan dataset, memasukkan analisis multimodal (seperti emoji atau gambar), serta menerapkan model *transformer* lain atau *fine-tuning* khusus untuk bahasa Indonesia agar dapat meningkatkan akurasi deteksi emosi dan relevansi topik secara lebih mendalam.

Daftar Pustaka

- [1] F. Li, J. Larimo, and L. C. Leonidou, "Social media in marketing research: Theoretical bases, methodological aspects, and thematic focus," *Psychol. Mark.*, vol. 40, no. 1, pp. 124–145, 2023, doi: 10.1002/mar.21746.
- [2] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *J. Inf. Technol. Case Appl. Res.*, vol. 25, no. 3, pp. 277–304, 2023, doi: 10.1080/15228053.2023.2233814.
- [3] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, pp. 1–20, 2021, doi: 10.1186/s40537-021-00459-1.
- [4] A. K. Sandhu, "Big Data with Cloud Computing: Discussions and Challenges," *Big Data Min. Anal.*, vol. 5, no. 1, pp. 32–40, 2022, doi: 10.26599/BDMA.2021.9020016.
- [5] V. Dogra *et al.*, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/1883698.
- [6] S. Mishra, P. Shukla, and R. Agarwal, "Review Article Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2022, 2022, doi: 10.1155/2022/1575365.
- [7] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, no. April 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.
- [8] J. Zhang, D. Wolfram, and F. Ma, "The impact of big data on research methods in information science," *Data Inf. Manag.*, vol. 7, no. 2, p. 100038, 2023, doi: 10.1016/j.dim.2023.100038.
- [9] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, and P. Rezaei-Hachesu, "Interoperability of heterogeneous health information systems: a systematic literature review," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, pp. 1–13, 2023, doi: 10.1186/s12911-023-02115-5.
- [10] S. Mondal, S. Das, and V. G. Vrana, "How to Bell the Cat? A Theoretical Review of Generative Artificial Intelligence towards Digital Disruption in All Walks of Life," *Technologies*, vol. 11, no. 2, 2023, doi: 10.3390/technologies11020044.
- [11] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00948-z.
- [12] M. Tayefi *et al.*, "Challenges and opportunities beyond structured data in analysis of electronic health records," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 13, no. 6, pp. 1–19, 2021, doi: 10.1002/wics.1549.
- [13] Q. Qiu, B. Wang, K. Ma, and Z. Xie, "Geological profile-text information association model of mineral exploration reports for fast analysis of geological content," *Ore Geol. Rev.*, vol. 153, no. December 2022, p. 105278, 2023, doi: 10.1016/j.oregeorev.2022.105278.
- [14] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, no. May, pp. 1–16, 2022, doi: 10.3389/fsoc.2022.886498.

- [15] A. M. Grisales, S. Robledo, and M. Zuluaga, "Topic Modeling: Perspectives From a Literature Review," *IEEE Access*, vol. 11, no. December 2022, pp. 4066–4078, 2023, doi: 10.1109/ACCESS.2022.3232939.
- [16] S. J. Weston, I. Shryock, R. Light, and P. A. Fisher, "Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial," *Adv. Methods Pract. Psychol. Sci.*, vol. 6, no. 2, 2023, doi: 10.1177/25152459231160105.
- [17] V. Taecharungroj, "'What Can ChatGPT Do?' Analyzing Early Reactions to the Innovative AI Chatbot on Twitter," *Big Data Cogn. Comput.*, vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010035.
- [18] M. S. González Canché, "Latent Code Identification (LACOID): A Machine Learning-Based Integrative Framework [and Open-Source Software] to Classify Big Textual Data, Rebuild Contextualized/Unaltered Meanings, and Avoid Aggregation Bias," *Int. J. Qual. Methods*, vol. 22, 2023, doi: 10.1177/16094069221144940.
- [19] U. Detthamrong *et al.*, "Topic Modeling Analytics of Digital Economy Research: Trends and Insights," *J. Scientometr. Res.*, vol. 13, no. 2, pp. 448–458, 2024, doi: 10.5530/jscires.13.2.35.
- [20] X. Wu, T. Nguyen, and A. T. Luu, "A survey on neural topic models: methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 57, no. 2, pp. 1–30, 2024, doi: 10.1007/s10462-023-10661-7.
- [21] M. T. Mohammed and O. F. Rashid, "Document retrieval using term frequency inverse sentence frequency weighting scheme," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 3, pp. 1478–1485, 2023, doi: 10.11591/ijeecs.v31.i3.pp1478-1485.
- [22] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access*, vol. 9, no. ML, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.
- [23] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Comput. Sci.*, vol. 8, pp. 1–28, 2022, doi: 10.7717/PEERJ-CS.914.
- [24] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports," *Math. Probl. Eng.*, vol. 2021, no. ii, 2021, doi: 10.1155/2021/6619088.
- [25] L. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/9285324.
- [26] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bull. Electr. Eng. Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021, doi: 10.11591/eei.v10i5.3157.
- [27] M. Y. Dhinora and E. Mailoa, "Analisa Tweet Mahasiswa untuk Deteksi Gejala Depresi dengan Penerapan Natural Language Processing," *J. Indones. Manaj. Inform. dan Komun.*, vol. 6, no. 2, pp. 1193–1211, 2025, doi: 10.63447/jimik.v6i2.1405.
- [28] Y. Inoue, "Multi-objective Optimization of BERTopic for Large-scale Scholarly Literature Analysis : Application to Autoregressive Models," pp. 1–11.
- [29] D. Hanny and B. Resch, "Clustering-Based Joint Topic-Sentiment Modeling of Social Media Data: A Neural Networks Approach," *Inf.*, vol. 15, no. 4, pp. 1–20, 2024, doi: 10.3390/info15040200.
- [30] C. Galli, M. T. Colangelo, M. Meleti, S. Guizzardi, and E. Calciolari, "Topic Analysis of the Literature Reveals the Research Structure: A Case Study in Periodontics," *Big Data Cogn. Comput.*, vol. 9, no. 1, pp. 1–27, 2025, doi: 10.3390/bdcc9010007.
- [31] G. Pranauskas, "Machine Learning Based Narrative Search in the Information Space," *Journals Vilniaus Univ.*, vol. 4, no. 1, 2025.
- [32] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU J. Sci. Technol.*, vol. 4,

- no. 1&2, pp. 27–33, 2022, doi: 10.56471/slujst.v4i.266.
- [33] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, “Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data,” *Heliyon*, vol. 10, no. 16, p. e35945, 2024, doi: 10.1016/j.heliyon.2024.e35945.
 - [34] F. Lan, “Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method,” *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/7923262.
 - [35] N. Aamir *et al.*, “Topic Modeling Empowered by a Deep Learning Framework Integrating BERTopic, XLM-R, and GPT,” *J. Comput. Biomed. Informatics*, vol. 8, no. 2, 2025, [Online]. Available: <https://doi.org/10.56979/802/2025>
 - [36] C. Galli, C. Cusano, M. Meleti, N. Donos, and E. Calciolari, “Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings,” *Metrics*, vol. 1, no. 1, p. 2, 2024, doi: 10.3390/metrics1010002.
 - [37] M. A. Mersha, M. Gameda Yigezu, and J. Kalita, “Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms,” *Procedia Comput. Sci.*, vol. 244, pp. 121–132, 2024, doi: 10.1016/j.procs.2024.10.185.
 - [38] A. Rejeb, K. Rejeb, S. Simske, and E. Süle, “Industry 5.0 research: an approach using co-word analysis and BERTopic modeling,” *Discov. Sustain.*, vol. 6, no. 1, 2025, doi: 10.1007/s43621-025-01252-3.
 - [39] M. Martinsuo and M. Huemann, “Reporting case studies for making an impact,” *Int. J. Proj. Manag.*, vol. 39, no. 8, pp. 827–833, 2021, doi: 10.1016/j.ijproman.2021.11.005.