BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

Pemodelan Topik pada Komunitas Ekspresi Emosi Negatif di Media Sosial X Menggunakan LDA

Rizal Muhammad Ramli, Chanifah Indah Ratnasari*

Fakultas Teknologi Industri, Program Studi Informatika, Universitas Islam Indonesia, Sleman, Indonesia
Email: [1]rizal.ramli@students.uii.ac.id, [2,*]chanifah.indah@uii.ac.id
Email Penulis Korespondensi: chanifah.indah@uii.ac.id

Abstrak−Penelitian ini bertujuan untuk memetakan struktur tematik percakapan dalam komunitas ekspresi emosi negatif di platform X, atau yang sering disebut sebagai "Komunitas MARAH MARAH". Permasalahan utama yang dikaji adalah bagaimana pola kemarahan kolektif terbentuk serta isu-isu apa saja yang mendominasi percakapan dalam komunitas tersebut. Untuk menjawab pertanyaan tersebut, penelitian menggunakan pendekatan *text mining* melalui serangkaian tahapan pemrosesan data teks, meliputi *scraping* data, *preprocessing*, normalisasi berbasis kamus, pembobotan *Term Frequency–Inverse Document Frequency* (TF-IDF), dan pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA). Sebanyak 75.032 *tweet* berhasil dikumpulkan, kemudian dibersihkan sehingga menghasilkan 38.956 data unik untuk dianalisis lebih lanjut. Pemodelan topik dilakukan dengan menguji sejumlah konfigurasi jumlah topik, dan *coherence score* tertinggi sebesar 0,5367 diperoleh pada model dengan tiga topik. Analisis lebih lanjut mengungkap tiga tema utama beserta proporsi dominasinya, yaitu keluhan personal dan curahan emosi sehari-hari (50,5%), ekspresi kemarahan langsung maupun generalisasi terhadap kelompok tertentu (35,8%), serta isu penipuan dan keamanan digital (13,7%). Temuan ini memberikan gambaran mengenai bagaimana kemarahan kolektif dikonstruksi, disebarkan, dan dimaknai dalam ruang percakapan daring. Penelitian ini diharapkan dapat menjadi dasar bagi kajian lanjutan mengenai emosi digital, dinamika komunitas daring, dan pemetaan isu sosial berbasis percakapan publik.

Kata Kunci: Analisis Teks; Kemarahan; Komunitas Daring; LDA; Media Sosial; Topic Modeling

Abstract−This study aims to map the thematic structure of conversations within a community of negative emotional expression on platform X, commonly referred to as the "Komunitas MARAH MARAH." The primary problem explored in this study is how collective anger is formed and which issues dominate the discourse within this community. To address this, the study employs a text mining approach through several stages of textual data processing, including data scraping, preprocessing, dictionary-based normalization, Term Frequency–Inverse Document Frequency (TF-IDF) weighting, and topic modeling using Latent Dirichlet Allocation (LDA). A total of 75,032 tweets were collected and subsequently cleaned, resulting in 38,956 unique entries for further analysis. Topic modeling was conducted by evaluating several topic configurations, with the highest coherence score of 0.5367 achieved using a three-topic model. Further analysis revealed three dominant themes along with their proportional distributions: personal complaints and everyday emotional expression (50.5%), direct anger or generalized expressions toward particular groups (35.8%), and issues related to fraud and digital security (13.7%). These findings illustrate how collective anger is constructed, disseminated, and interpreted within online conversational spaces. This study is expected to serve as a foundation for further research on digital emotion, online community dynamics, and social issue mapping through public discourse.

Keywords: Text Analysis; Anger; Online Community; LDA; Social Media; Topic Modeling

1. PENDAHULUAN

Media sosial telah berkembang menjadi ruang publik digital yang memungkinkan masyarakat mengekspresikan opini, emosi, dan respons terhadap berbagai persoalan sosial secara cepat, masif, dan terbuka. Platform X (sebelumnya Twitter) merupakan salah satu media sosial berbasis teks pendek yang paling dominan digunakan untuk diskusi publik, penyebaran informasi, hingga pembentukan persepsi kolektif terhadap isu-isu aktual [1], [2]. Dalam konteks komunikasi digital tersebut, ekspresi emosi, khususnya emosi negatif seperti kemarahan, sering kali mendominasi percakapan dan dapat mencerminkan dinamika sosial yang sedang berlangsung, baik terkait layanan publik, hubungan interpersonal, maupun tekanan sosial-ekonomi [3], [4]. Fenomena tersebut menunjukkan bahwa ruang digital tidak hanya menjadi sarana komunikasi, tetapi juga menjadi wadah representasi kondisi psikologis sekaligus potret respons masyarakat terhadap perubahan sosial.

Penelitian sebelumnya menunjukkan bahwa konten bermuatan emosi kuat, khususnya kemarahan, memiliki kecenderungan untuk menyebar lebih cepat dibandingkan emosi positif. Sifatnya yang reaktif dan provokatif membuat pengguna lebih terdorong untuk merespons, membagikan, atau memperdebatkan konten tersebut [5]. Selain itu, algoritma media sosial secara inheren dirancang untuk meningkatkan *engagement*, sehingga cenderung memperkuat distribusi konten emosional dan memperluas jangkauannya [6]. Akibatnya, percakapan bermuatan negatif tidak hanya menyebar lebih cepat, tetapi juga berpotensi membentuk opini kolektif yang memengaruhi persepsi publik terhadap isu tertentu.

Dalam konteks ini, muncul berbagai komunitas digital yang berfungsi sebagai kanal ekspresi emosional. Salah satunya adalah komunitas ekspresi emosi negatif di platform X, atau yang disebut oleh pengguna sebagai "Komunitas MARAH MARAH". Komunitas ini digunakan sebagai ruang oleh pengguna untuk mengekspresikan frustrasi, keluhan, dan kemarahan. Dalam penelitian ini, istilah tersebut dipertahankan sebagai istilah teknis dan diperlakukan sebagai konstruksi sosial yang menggambarkan fenomena komunitas digital berbasis ekspresi emosi negatif, sehingga dapat dianalisis dalam kerangka akademik. Fenomena serupa juga ditemukan dalam *venting communities* di Reddit dan *ranting threads* di platform lain, yang menunjukkan adanya pola interaksi emosional khas dan konsisten di berbagai konteks budaya [7], [8].

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

Kajian akademik mengenai ekspresi emosi negatif di media sosial telah banyak dibahas dalam konteks politik [9], kebijakan publik [10], serta kesehatan mental [11]. Meskipun demikian, penelitian yang secara khusus memotret komunitas yang memang berfokus pada ekspresi kemarahan masih terbatas, terutama di Indonesia. Sebagian besar studi terdahulu menitikberatkan pada *sentiment analysis* atau klasifikasi emosi, sedangkan pemetaan struktur tematik atau *topical structure* dari kemarahan kolektif belum mendapatkan perhatian yang memadai [12], [13]. Analisis sentimen memang memberikan gambaran polaritas, tetapi tidak mampu menjelaskan konteks, ragam isu, atau struktur diskursus yang melatarbelakangi ekspresi kemarahan.

Sementara itu, pendekatan *topic modeling*, khususnya *Latent Dirichlet Allocation* (LDA), telah terbukti efektif dalam mengidentifikasi pola tematik pada data teks berskala besar, termasuk data teks pendek seperti *tweet* [14], [15]. Berbagai penelitian juga mengeksplorasi metode pembanding seperti *Non-Negative Matrix Factorization* (NMF) dan BERTopic untuk ekstraksi tema laten. Namun, LDA masih dianggap lebih stabil, lebih mudah direplikasi, dan lebih interpretatif pada konteks analisis teks pendek dengan jumlah dokumen besar. Hal ini menjadikan LDA metode yang relevan untuk memetakan dinamika ekspresi emosi dalam komunitas daring yang berkembang secara organik.

Gap penelitian dalam topik ini muncul dari beberapa aspek. Pertama, belum adanya kajian yang secara khusus memetakan topik kemarahan dalam komunitas digital yang didedikasikan untuk ekspresi emosi negatif, bukan percakapan umum atau isu spesifik. Kedua, penelitian yang ada mengenai ekspresi emosi daring cenderung bersifat deskriptif dan belum menghubungkan temuan tematik dengan konteks sosial yang lebih luas, terutama terkait pembentukan kemarahan kolektif [16], [17]. Ketiga, sebagian penelitian masih menggunakan pendekatan *supervised learning* sehingga hanya menghasilkan klasifikasi polaritas tanpa menggali struktur diskursus yang lebih kompleks. Dalam konteks tersebut, LDA memberikan kemampuan untuk mengidentifikasi isu-isu utama secara tidak terawasi (*unsupervised*) sehingga temuan yang diperoleh lebih kaya dan mendalam.

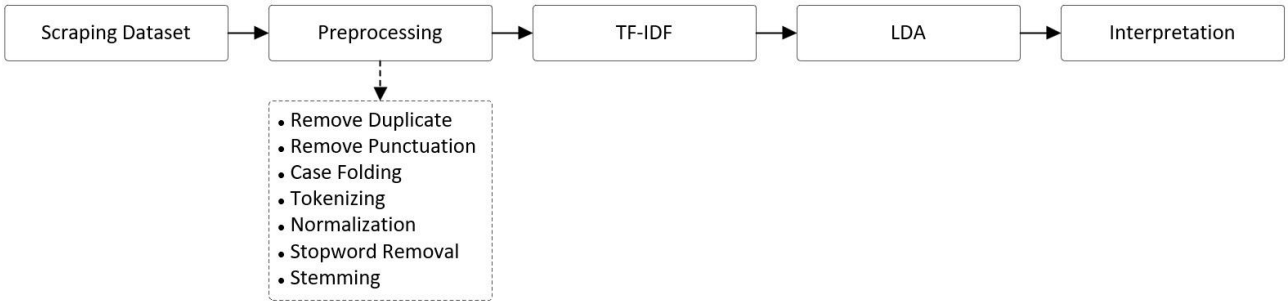
Berdasarkan gap tersebut, penelitian ini memandang penting untuk melakukan pemetaan tematik secara komprehensif terhadap ekspresi kemarahan kolektif dalam komunitas ekspresi emosi negatif di platform X. Tujuan penelitian ini adalah mengidentifikasi dan memodelkan topik-topik utama yang muncul dalam komunitas tersebut dengan menggunakan metode LDA. Cakupan penelitian meliputi: (1) pengumpulan dan pengolahan data teks, (2) penerapan pemodelan topik untuk mengekstraksi tema laten, dan (3) interpretasi hasil topik dalam konteks sosial yang lebih luas.

Penelitian ini diharapkan dapat memberikan kontribusi teoretis bagi literatur mengenai sosiologi digital, analisis emosi kolektif, dan studi percakapan daring di Indonesia [18], [19]. Selain itu, hasilnya juga dapat menjadi dasar pemahaman bagi pemangku kebijakan dan analis media sosial untuk mengidentifikasi isu-isu yang memicu ketidakpuasan publik serta menilai bagaimana emosi negatif menyebar dan membentuk dinamika diskursus dalam ruang digital.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini terdiri dari beberapa tahapan, yaitu pengumpulan data, *preprocessing*, representasi teks, pemodelan topik, dan interpretasi hasil. Tahapan ini mengacu pada pendekatan umum dalam *text mining* dan analisis media sosial [20]. Alur lengkap tahapan penelitian ditunjukkan pada Gambar 1, yang menampilkan urutan proses mulai dari *scraping* hingga interpretasi topik.



Gambar 1. Tahapan Penelitian

Sebagaimana diperlihatkan pada Gambar 1, tahap pertama adalah pengumpulan data dari komunitas ekspresi emosi negatif ("marah-marah") di platform X menggunakan teknik *scraping*. Tahap ini menghasilkan data *tweet* mentah yang berisi keluhan, ujaran emosional, dan ungkapan kemarahan pengguna. Tahap kedua adalah *preprocessing*, yaitu tahap pembersihan data yang dilakukan melalui beberapa langkah, meliputi penghapusan duplikasi, penghilangan tanda baca, *case folding*, tokenisasi, normalisasi teks, *stopword removal* menggunakan kamus Sastrawi, dan *stemming*. Langkah-langkah tersebut mengacu pada praktik pemrosesan bahasa alami untuk bahasa Indonesia sebagaimana direkomendasikan dalam penelitian NLP (*natural language processing*) bahasa Indonesia [21], [22]. *Preprocessing* ini dilakukan untuk memastikan teks berada dalam bentuk yang sesuai untuk proses penambangan teks [23], [24], dalam hal ini penambangan topik.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

Tahap berikutnya adalah representasi teks menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai salah satu metode umum dalam pembobotan kata pada dokumen [25]. Pembobotan ini berfungsi memberikan nilai berbeda bagi kata yang memiliki tingkat kekhususan lebih tinggi sehingga dapat meningkatkan kualitas identifikasi topik.

Selanjutnya, pemodelan topik dilakukan menggunakan *Latent Dirichlet Allocation* (LDA) yang diperkenalkan oleh Blei et al. [14]. LDA dipilih karena mampu mengekstraksi struktur laten dari teks berjumlah besar dan telah banyak digunakan dalam penelitian analisis media sosial. Pada penelitian ini, beberapa parameter seperti *num_topics*, *passes*, *iterations*, *alpha*, dan *random_state* diatur sesuai eksperimen untuk memperoleh nilai *coherence score* terbaik. *Coherence score* merupakan salah satu metrik untuk mengevaluasi kualitas model sebagaimana telah divalidasi dalam penelitian-penelitian sebelumnya [26]. Terakhir, hasil topik dievaluasi melalui analisis tema dan interpretasi konteks sosial.

2.2 *Preprocessing* Data dan Pembobotan TF-IDF

Tahapan *preprocessing* dilakukan untuk mengubah data mentah menjadi bentuk teks yang bersih dan siap diproses oleh model. Proses ini meliputi:

a. Penghapusan duplikasi untuk menghapus *tweet* yang sama menggunakan fitur *Remove Duplicate* pada Excel.
b. *Remove punctuation* untuk menghilangkan semua karakter tanda baca, angka, dan simbol khusus yang tidak memiliki makna semantik pada konteks topik.
c. *Case folding* untuk menyeragamkan teks menjadi format huruf kecil (*lowercase*) agar konsisten.
d. Tokenisasi untuk memecah teks menjadi unit kata.
e. Normalisasi untuk memperbaiki kata-kata tidak baku, slang, atau singkatan yang disesuaikan ke bentuk baku. Mengingat karakteristik data media sosial yang penuh dengan teks yang tidak baku, tahap ini menjadi sangat krusial. Proses normalisasi dilakukan menggunakan kamus konversi internal yang disusun secara semi otomatis. Kamus ini memetakan 4.000+ pasangan kata slang–kata baku, termasuk variasi umum seperti "gw→saya", "bgt→banget", "kzl→kesal", dan "anjir→anjing". Proses normalisasi dilakukan dengan pencocokan langsung pada token untuk mengganti kata tidak baku menjadi kata standar bahasa Indonesia.
f. *Stopwords removal* untuk menghilangkan kata-kata umum yang tidak memberikan kontribusi pada pembahasan topik menggunakan daftar *stopwords* Sastrawi.
g. *Stemming* untuk mengembalikan setiap kata ke bentuk dasarnya, dalam penelitian ini menggunakan *library* Sastrawi.

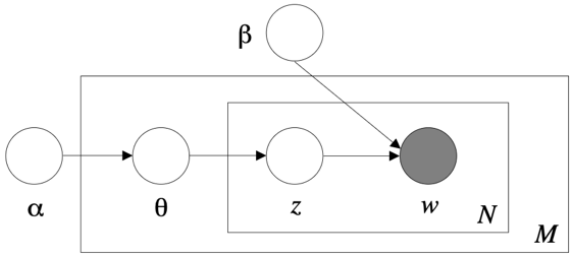
Dalam *dataset* ini, sebagian *tweet* menggunakan campuran bahasa Indonesia dan bahasa Inggris, yang merupakan karakteristik umum percakapan daring. Namun, penelitian ini belum menerapkan *language detection* untuk memisahkan kedua bahasa tersebut. Hal ini menjadi catatan pekerjaan rumah penelitian lanjutan, sebaiknya ditambahkan mekanisme deteksi bahasa atau *filtering* untuk bahasa asing agar model lebih konsisten.

Setelah *preprocessing*, representasi dokumen menggunakan TF-IDF dilakukan untuk menghasilkan bobot kata. TF-IDF merupakan metode yang telah banyak digunakan pada analisis dokumen pendek, termasuk *tweet* [27]. Hasil pembobotan kemudian menjadi masukan untuk proses pemodelan topik LDA.

2.3 *Topic Modeling* menggunakan LDA

LDA bekerja dengan mengasumsikan bahwa setiap dokumen merupakan campuran dari beberapa topik, dan setiap topik tersusun dari sejumlah kata kunci dengan probabilitas tertentu [14]. Model ini menghasilkan dua keluaran utama: (1) daftar kata kunci pada tiap topik, dan (2) distribusi topik pada tiap dokumen. Dalam penelitian ini, kualitas model dinilai menggunakan *coherence score*, sebagaimana direkomendasikan oleh Röder et al. [26].

Dalam penelitian ini, model LDA diimplementasikan menggunakan *library* Gensim pada Python. Dengan menerima korpus dalam format *bag-of-words* sebagai input, model LDA akan mengidentifikasi distribusi kata untuk setiap topik dan distribusi topik untuk setiap dokumen. Jumlah topik yang akan diekstraksi ditentukan sebagai parameter input untuk model.



Gambar 2. Visualisasi Grafis Model LDA

Pada Gambar 2, diperlihatkan representasi *plate notation* dari model LDA [28]. Gambar tersebut menampilkan struktur probabilistik LDA, di mana *plate notation* menggambarkan pengulangan unit dokumen dan kata di dalam dokumen. Pelat bagian luar merepresentasikan kumpulan dokumen (M), menandakan bahwa proses pemilihan topik dan kata berlaku untuk setiap dokumen dalam korpus. Pelat bagian dalam menunjukkan pengulangan proses pemilihan topik dan kata untuk setiap token (kata ke-n) dalam sebuah dokumen. Adapun penjelasan variabel pada pelat adalah sebagai berikut.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

α : hiperparameter distribusi topik–dokumen
β : hiperparameter distribusi kata–topik
θ_d : distribusi topik untuk dokumen d
z_n : topik yang dipilih untuk token ke-n
w_n : token/word yang diobservasi
N : jumlah kata dalam dokumen
M : jumlah dokumen dalam korpus

Penelitian ini menggunakan *coherence score* sebagai metrik utama untuk mengevaluasi kualitas topik. Evaluasi dilakukan menggunakan modul `CoherenceModel` dari *library* Gensim. Proses pelatihan model LDA (`LdaModel`) dijalankan secara iteratif dengan parameter yang telah didefinisikan seperti, `corpus`, `dictionary`, `num_topic`, `random_state`, `iterations`, `passes`, dan `alpha` untuk memperoleh model topik yang paling optimal. Proses perhitungan tersebut mengikuti formulasi probabilistik LDA sebagaimana ditunjukkan pada persamaan (1), menggambarkan proses generatif dalam pemilihan topik dan kata pada tiap dokumen.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (1)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dataset penelitian dikumpulkan dari komunitas "Komunitas MARAH MARAH" pada platform X (https://x.com/i/communities/1562271278744354816). Pengambilan data dilakukan menggunakan *library* Selenium pada rentang waktu 21 Februari hingga 31 Juli 2025. Seluruh proses *scraping* dilakukan secara bertahap dan menghasilkan data mentah sebanyak 75.032 *tweet*.

Sebelum dianalisis, data mentah tersebut harus melalui proses pembersihan berupa deteksi dan penghapusan data duplikat. Hasil proses deteksi duplikasi ditunjukkan pada Gambar 3, yang memperlihatkan jumlah *tweet* yang duplikat dan unik. Pada gambar tersebut, dari total 75.032 *tweet*, terlihat bahwa sebanyak 36.076 *tweet* teridentifikasi sebagai duplikasi dan dihapus, sehingga menyisakan 38.956 *tweet* unik yang digunakan untuk tahapan analisis berikutnya.

2025-02-20T17:04:04.000Z	@notsopurplesky	BACOTTT BGTT KONTOLLLLL PENYEPONG WOWO
2025-02-20T17:05:19.000Z	@steroxoo	BABI JADI BACKBURNER DIPERTEMANAN GA ENAK BGT ASUU. MENDING JADI BACKBURNER SAMA LAWAN JENIS ANJINGGGGG
2025-02-20T17:06:14.000Z	@agunbuhori	Heran gw ada orang minjem uang puluhan juta janji bayar dalam waktu dekat tapi dia sendiri kerjaannya gak jelas, mana keluarga sendiri lagi. Gak dikasih malah ngamok
2025-02-20T17:06:39.000Z	@weneedmorem0ney	Kata gua sih anjing, gua minta tolong dan ngomong baik-baik tapi lu so begaya kaya orang paling superior gatau diri anjing.
2025-02-20T17:07:16.000Z	@namsrchive	EH ANJING TOLOL FAEDAHNYA NGEFAKER TUH APASIIIII?????? lu saking insecurenya kah sampe pake pake muka orang??? monyett banget muka tmn gue difakerin udah gitu dijadiin bisnis. tai kuda dikasi nyawa begini nih kelakuannya
2025-02-20T17:09:02.000Z	@KoreAquila	Ada gak yang kelamaan nganggur di rumah malah menciptakan toxic relation sama keluarga? Kesannya semua salah. Gua baru 2 minggu usai ukom dan baru aja pengumuman lulus, yudis dan str aja belom dpt. Udah apply sana sini, belom ada panggilan. Sesusah itu cari kerja?
2025-02-20T17:10:08.000Z	@cloudiesgo	DOAIN GUE LAGI RIBUT SAMA BANYAK KOMENAN DI VT INI, CUMA GARA" GA HAFAL LAGU DARAH JUANG MALAH BILANG DEMO JUGA GA RELEVAN GOBLOK
2025-02-20T17:11:30.000Z	@ayselew	anjing capek bgt punya badan kurus bb dr 41 turun ke 40 turun ke 39 abis itu turun ke 38
2025-02-20T17:11:40.000Z	@necromanticcs	Normalize bully yang punya pacar parcok
2025-02-20T17:12:04.000Z	@madeforbucky	MAU MARAH BGT tp ujungnya nangis. Aku baru beli set buku yg menurutku mahal tp dapetnya beginii T___T bolongnya nembus sampe ke boxnya huhuhu. Menurut kalian aku harus komplen ke sellernya atau ke ekspedisinya dulu ya?:((

Gambar 3. Contoh Data Unik

3.2 *Preprocessing*

Tahap *preprocessing* dilakukan untuk membersihkan dan menstandarkan data teks sehingga siap digunakan dalam proses ekstraksi fitur dan pemodelan topik. Tahapan yang diterapkan meliputi: (1) *remove punctuation*, (2) *case folding*, (3) *tokenizing*, (4) *normalization*, (5) *stopwords removal*, dan (6) *stemming*. Setiap tahap dijelaskan menggunakan contoh data sebagaimana ditampilkan pada Tabel 1 hingga Tabel 6.

Remove punctuation dilakukan untuk mengeliminasi karakter tanda baca, angka, dan simbol khusus yang dianggap sebagai *noise* dari korpus teks. Contoh dari proses *remove punctuation* ditampilkan pada Tabel 1, yang menunjukkan bagaimana karakter non-alfabet seperti tanda baca, angka, dan simbol dihilangkan untuk mengurangi *noise* pada data.

Tabel 1. Contoh dari *Remove Punctuation*

Data Mentah	*Remove Punctuation*
Udah gatau lagi mau gimana. Cuma mau mereka" itu dikasi sanksi sosial aja. Cuma di X harapan gw	Udah gatau lagi mau gimana Cuma mau mereka itu dikasi sanksi sosial aja Cuma di X harapan gw

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

Seperti terlihat pada Tabel 1, teks yang awalnya memiliki tanda kutip dan simbol berubah menjadi bentuk yang lebih bersih sehingga memudahkan proses *preprocessing* selanjutnya.

Tahap berikutnya adalah *case folding*. Teks hasil dari tahap *remove punctuation* kemudian menjadi masukan untuk tahap *case folding*. Pada tahap ini, seluruh karakter dikonversi menjadi huruf kecil (*lowercase*). Contoh hasil transformasi ini dapat dilihat pada Tabel 2.

Tabel 2. Contoh dari *Case Folding*

Remove Punctuation	*Case Folding*
Udah gatau lagi mau gimana Cuma mau mereka itu dikasi sanksi sosial aja Cuma di X harapan gw	udah gatau lagi mau gimana cuma mau mereka itu dikasi sanksi sosial aja cuma di x harapan gw

Tabel 2 menunjukkan bahwa seluruh token kini berada dalam format huruf kecil sehingga konsistensi korpus lebih terjaga. Setelah teks menjadi seragam huruf kecil semua, struktur data diubah melalui proses *tokenizing*. Seperti yang diilustrasikan pada Tabel 3, teks yang sebelumnya utuh kini dipecah menjadi daftar token (unit-unit kata) individual. Ini merupakan format fundamental yang diperlukan untuk analisis lebih lanjut.

Tabel 3. Contoh dari *Tokenizing*

Case Folding	*Tokenizing*
udah gatau lagi mau gimana cuma mau mereka itu dikasi sanksi sosial aja cuma di x harapan gw	[‘udah’, ‘gatau’, ‘lagi’, ‘mau’, ‘gimana’, ‘cuma’, ‘mau’, ‘mereka’, ‘itu’, ‘dikasi’, ‘sanksi’, ‘sosial’, ‘aja’, ‘cuma’, ‘di’, ‘x’, ‘harapan’, ‘gw’]

Daftar token dari Tabel 3 selanjutnya distandardisasi melalui proses *normalization*. Mengingat data berasal dari media sosial, tahap ini sangat krusial untuk mengubah token-token tidak baku, slang, atau singkatan menjadi kata baku. Hasil pada Tabel 4 menunjukkan bahwa kata tidak baku seperti '*gw*' berhasil dikonversi menjadi bentuk baku ('*saya*'), sehingga mengurangi variasi kata yang tidak diperlukan dalam analisis.

Tabel 4. Contoh dari Normalisasi

Tokenizing	Normalisasi
['udah', 'gatau', 'lagi', 'mau', 'gimana', 'cuma', 'mau', 'mereka', 'itu', 'dikasi', 'sanksi', 'sosial', 'aja', 'cuma', 'di', 'x', 'harapan', 'gw']	sudah gatau lagi mau gimana hanya mau mereka itu dikasi sanksi sosial saja hanya di x harapan saya

Setelah diperoleh korpus teks yang sudah baku melalui tahap normalisasi, dilakukan *stopwords removal* untuk menghilangkan kata-kata umum yang tidak berkontribusi terhadap pembentukan topik. Contohnya ditampilkan pada Tabel 5.

Tabel 5. Contoh dari *Stopwords Removal*

Normalisasi	*Stopwords Removal*
sudah gatau lagi mau gimana hanya mau mereka itu dikasi sanksi sosial saja hanya di x harapan saya	gatau mau gimana mau dikasi sanksi sosial x harapan

Tabel 5 menunjukkan bagaimana kata-kata seperti “di”, “itu”, dan “hanya” dihapus sehingga teks menjadi lebih padat makna. Sebagai tahap akhir *preprocessing* dalam penelitian ini, korpus teks yang tersisa dikenai proses *stemming*. Tabel 6 memperlihatkan kata-kata yang telah dilakukan *stemming*, di mana setiap token diubah menjadi bentuk kata dasarnya. *Output* dari tahap *stemming* ini berupa token-token dasar yang menjadi masukan untuk tahap pembobotan TF-IDF.

Tabel 6. Contoh dari *Stemming*

Stopwords Removal	*Stemming*
gatau mau gimana mau dikasi sanksi sosial x harapan	gatau mau gimana mau kasi sanksi sosial x harap

3.3 TF-IDF

Setelah melewati tahap *preprocessing*, dokumen yang berbentuk daftar token dikonversi menjadi representasi numerik. Pada tahap awal, data diubah menjadi format *bag-of-words* (BoW) menggunakan fungsi doc2bow untuk mendapatkan frekuensi kemunculan kata (*term frequency*/TF). Namun, penggunaan frekuensi murni memiliki kelemahan, kata-kata yang sering muncul namun tidak spesifik (misalnya kata "orang" atau "hari" yang mungkin lolos dari *stopwords*) akan mendominasi model, sehingga topik menjadi bias dan sulit dibedakan.

Untuk mengatasi hal tersebut, penelitian ini menerapkan skema pembobotan *Term Frequency–Inverse Document Frequency* (TF-IDF) menggunakan modul TfidfModel. Transformasi ini berfungsi untuk menyeimbangkan bobot kata. Algoritma secara otomatis memberikan "hukuman" (bobot rendah) pada kata yang muncul di hampir seluruh dokumen

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

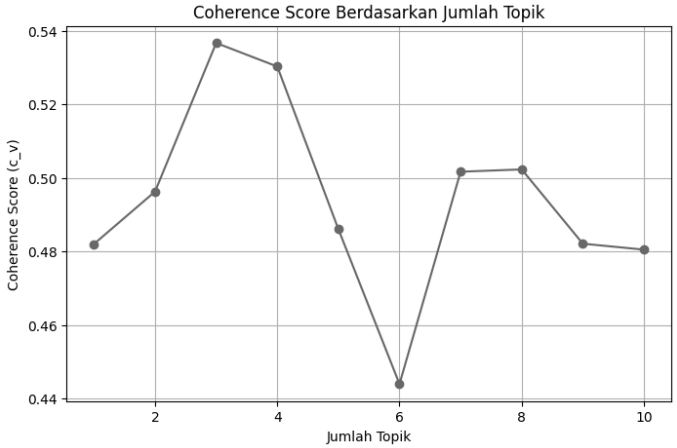
karena dianggap tidak memiliki daya pembeda, dan sebaliknya memberikan bobot tinggi pada kata-kata yang jarang muncul secara global namun dominan dalam dokumen tertentu.

Penggunaan TF-IDF memiliki dampak langsung terhadap kualitas representasi dokumen pada pemodelan topik. Skema ini menurunkan bobot kata-kata yang sering muncul tetapi tidak memiliki makna diskriminatif (misalnya "mau", "aja", "banget"), sekaligus memperbesar bobot kata-kata yang lebih spesifik terhadap konteks kemarahan atau keluhan (seperti "tipu", "anjing", atau "bantu"). Hasil matriks pembobotan inilah yang menjadi representasi fitur final untuk melatih model LDA, memastikan bahwa pembentukan topik didasarkan pada kekhasan semantik (*semantic distinctiveness*), bukan sekadar frekuensi leksikal.

3.4 *Topic Modeling*

Pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA) diterapkan untuk mengidentifikasi struktur topik laten yang tersembunyi dalam *dataset*. Penentuan jumlah topik (*num_topic*) yang optimal dilakukan menggunakan *coherence score* (c_v). Sangat penting untuk menilai seberapa koheren kumpulan kata yang dihasilkan oleh model dan seberapa mudah topik tersebut dapat diinterpretasikan secara manusiawi. Skor yang lebih tinggi mengindikasikan interpretabilitas yang lebih baik.

Untuk mendapatkan konfigurasi model terbaik, dilakukan serangkaian pengujian dengan memvariasikan jumlah topik dari 1 hingga 10. Tren perubahan nilai koherensi dari setiap skenario jumlah topik divisualisasikan dalam grafik garis sebagaimana ditampilkan pada Gambar 4.



Gambar 4. *Coherence Score* Berdasarkan Jumlah Topik

Pola kenaikan dan penurunan *coherence score* untuk tiap jumlah topik dapat diamati pada Gambar 4, yang menunjukkan performa terbaik terjadi pada jumlah 3 topik. Nilai lengkap *coherence score* untuk seluruh konfigurasi jumlah topik ditampilkan pada Tabel 7.

Tabel 7. *Conherence Value*

Jumlah Topik	*Conherence Score*
1	0,4820
2	0,4963
3	0,5367
4	0,5303
5	0,4860
6	0,4439
7	0,5017
8	0,5023
9	0,4821
10	0,4805

Berdasarkan Tabel 7, nilai *coherence score* tertinggi adalah 0,5367, dicapai saat model dikonfigurasi menggunakan 3 topik. Nilai ini lebih tinggi dibandingkan konfigurasi lainnya, seperti 2 topik (0,4963) atau 4 topik (0,5303). Merujuk pada temuan objektif ini, maka jumlah 3 topik ditetapkan sebagai parameter final (*num_topic*) untuk model LDA, karena dianggap mampu menghasilkan pemisahan tema yang paling stabil dan bermakna.

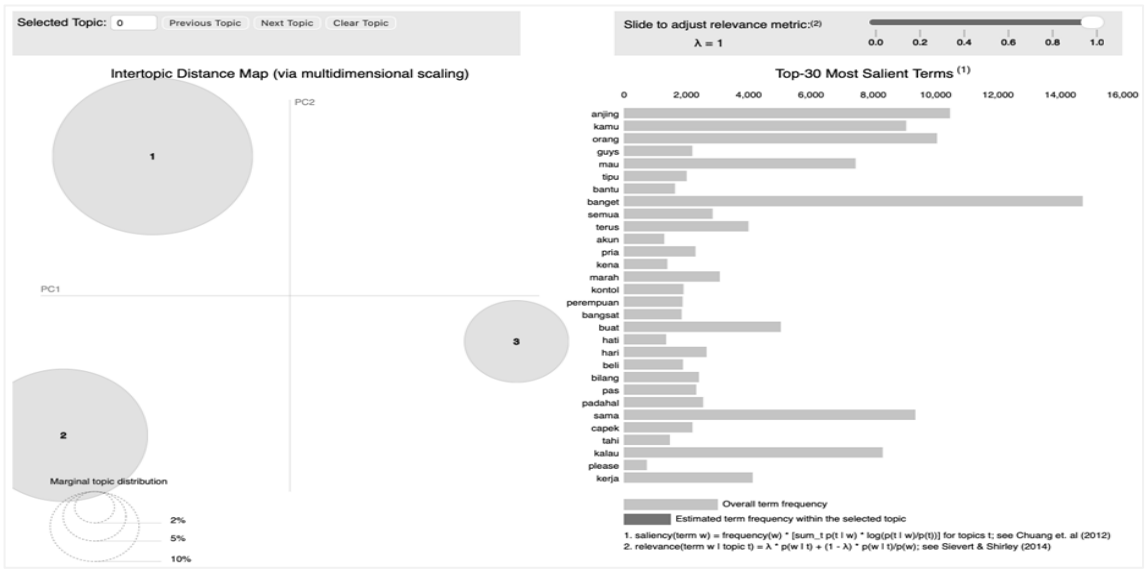
3.5 Visualisasi dan Interpretasi Topik

Untuk menganalisis dan menginterpretasi hasil dari 3 topik optimal tersebut, penelitian ini memanfaatkan *tool* visualisasi interaktif *pyLDAvis*. Visualisasi ini menyajikan dua bagian utama: pertama, bagian kiri adalah pemetaan topik dalam ruang dua dimensi yang direpresentasikan sebagai gelembung. Ukuran setiap gelembung menandakan frekuensi

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

kemunculan topik tersebut dalam keseluruhan *dataset*, sementara jarak antar gelembung menunjukkan tingkat korelasi semantik. Gelembung yang berjauhan atau tidak beririsan mengindikasikan topik yang berbeda secara tematik. Kedua, bagian kanan adalah diagram batang horizontal yang menampilkan 30 *term* (kata) paling relevan yang mewakili topik, di mana panjang diagram batang merepresentasikan frekuensi kata tersebut dalam konteks topik yang bersangkutan.

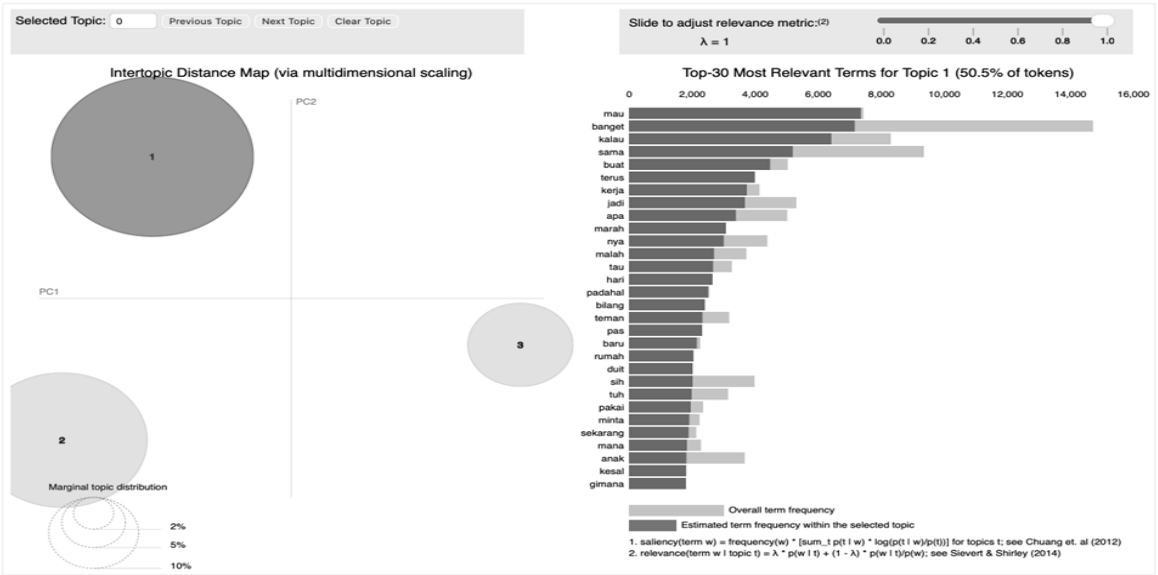
Sebagai langkah awal, peta jarak antar-topik (*intertopic distance map*) divisualisasikan untuk melihat seberapa baik model memisahkan setiap klaster secara global. Visualisasi keseluruhan model LDA yang menunjukkan sebaran topik pada "Komunitas MARAH MARAH" ditampilkan pada Gambar 5.



Gambar 5. Visualisasi Topik Global

Berdasarkan Gambar 5, terlihat bahwa model berhasil membentuk tiga klaster (gelembung) yang terpisah satu sama lain dalam kuadran yang berbeda. Hal ini mengindikasikan bahwa ketiga topik tersebut memiliki ciri khas semantik yang unik dan tidak tumpang tindih (*overlapping*). Pada gambar tersebut dapat dilihat bahwa topik ke-1 memiliki gelembung terbesar, dengan proporsi 50,5% dari total percakapan, diikuti topik ke-2 dan ke-3 dengan masing-masing proporsi sebesar 35,8% dan 13,7%.

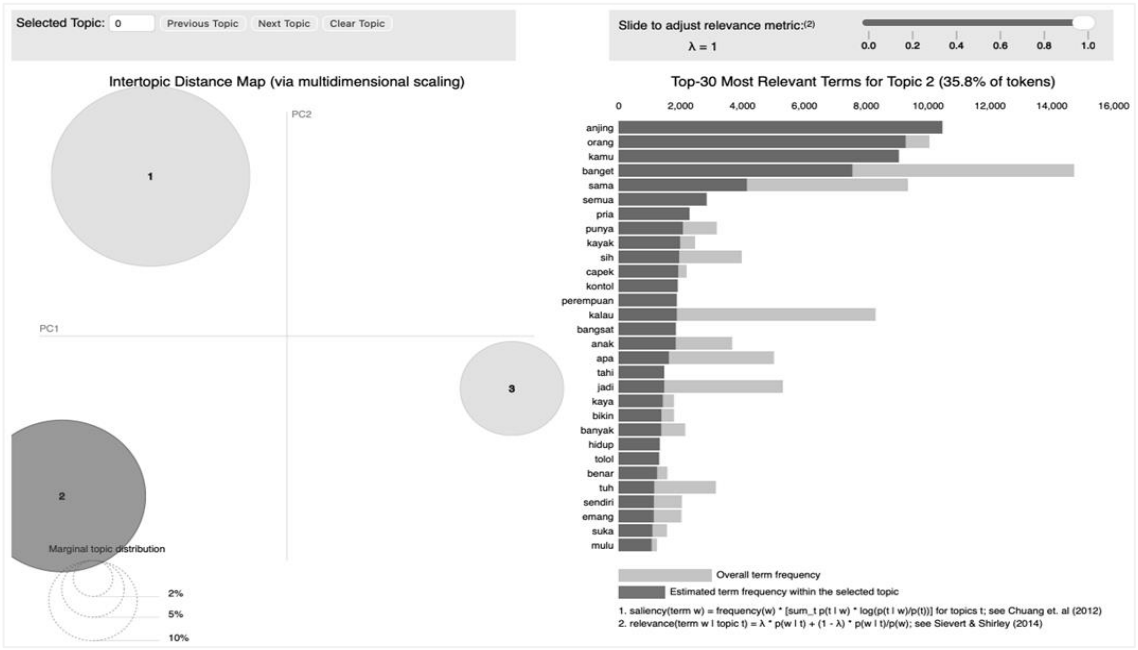
Selanjutnya, analisis dilakukan secara mendalam pada topik pertama yang memiliki proporsi token terbesar. Visualisasi pada Gambar 6 memperlihatkan distribusi kata kunci (*term frequency*) atau daftar kata paling representatif yang membentuk topik ke-1.



Gambar 6. Visualisasi Topik ke-1

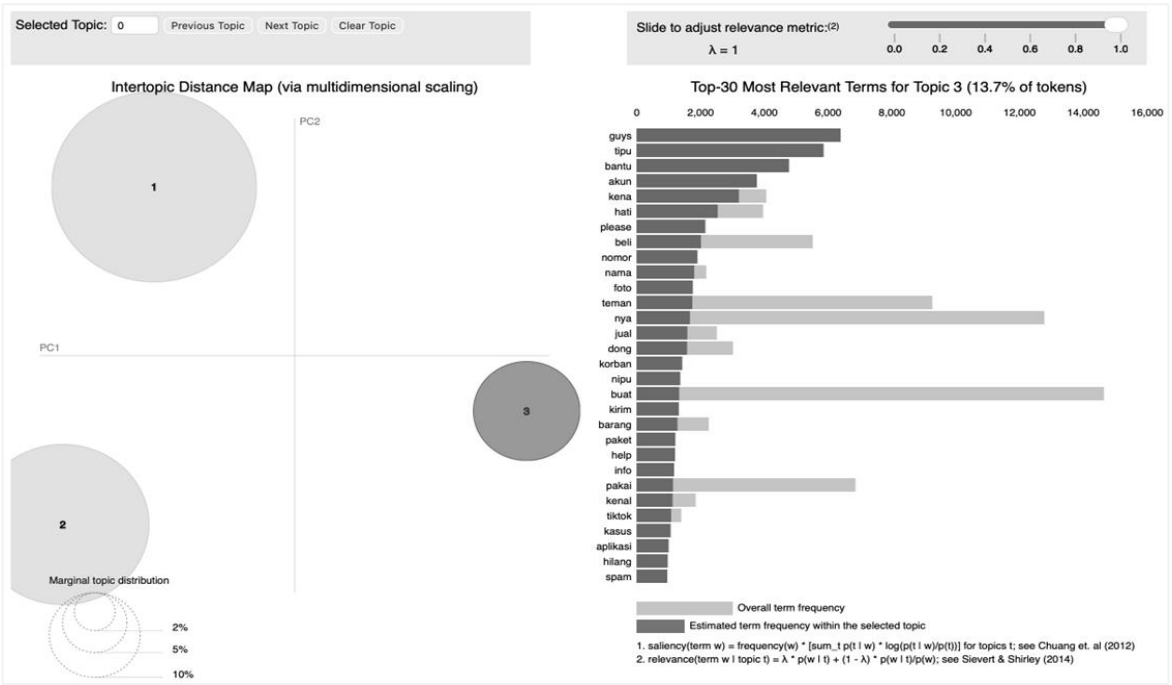
Berdasarkan Gambar 6, representasi kata-kata dengan probabilitas tertinggi pada topik ke-1 adalah: 0,022*"mau" + 0,021*"banget" + 0,019*"kalau" + 0,015*"sama" + 0,013*"buat" + 0,012*"terus" + 0,011*"kerja" + 0,011*"jadi" + 0,010*"apa" + 0,009*"marah". *Score* tertinggi terdapat pada kata “mau”, kemudian kata “banget”.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877



Gambar 7. Visualisasi Topik ke-2

Visualisasi dari topik ke-2 disajikan pada Gambar 7. Representasi kata-kata dengan probabilitas tertinggi pada topik ke-2 adalah: 0,043*"anjing" + 0,038*"orang" + 0,038*"kamu" + 0,031*"banget" + 0,017*"sama" + 0,012*"semua" + 0,010*"pria" + 0,009*"punya" + 0,008*"kayak" + 0,008*"sih". *Score* tertinggi terdapat pada kata “anjing”, kemudian kata “orang” dan “kamu”.



Gambar 8. Visualisasi Topik ke-3

Visualisasi dari topik ke-3 disajikan pada Gambar 8. Representasi kata-kata dengan probabilitas tertinggi pada topik ini adalah: 0,024*"guys" + 0,022*"tipu" + 0,018*"bantu" + 0,014*"akun" + 0,012*"kena" + 0,009*"hati" + 0,008*"please" + 0,008*"beli" + 0,007*"nomor" + 0,007*"nama". *Score* tertinggi terdapat pada kata “guys”, kemudian kata “tipu”.

Setelah model LDA berhasil mengidentifikasi klaster-klaster topik dari *corpus* data, langkah selanjutnya adalah interpretasi. Hasil dari LDA hanya menyajikan kumpulan kata kunci probabilistik untuk setiap topik, yang belum memiliki makna eksplisit. Oleh karena itu, untuk memahami topik secara lebih mendalam, dilakukan analisis interpretasi topik secara detail. Proses analisis kualitatif ini bertujuan untuk 'menerjemahkan' kumpulan *term* tersebut menjadi sebuah tema yang koheren dan substantif. Langkah ini krusial agar orang awam dapat dengan mudah memahami konteks dan

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 224-234
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.877

substansi dari isu-isu tersembunyi yang menjadi fokus utama dalam *dataset* komunitas marah-marah ini. Interpretasi topik dijabarkan dalam Tabel 8.

Tabel 8. Interpretasi Topik

Topik	Kata Kunci	Penjelasan
1	0,022*"mau" + 0,021*"banget" + 0,019*"kalau" + 0,015*"sama" + 0,013*"buat" + 0,012*"terus" + 0,011*"kerja" + 0,011*"jadi" + 0,010*"apa" + 0,009*"marah"	Pola bicara umum yang mengekspresikan keluhan atau curhat, seringkali berisi kata-kata emosional ("banget", "marah") terkait suatu situasi personal, pekerjaan, emosi sehari-hari ("kerja", "terus"-menerus) atau keinginan ("mau").
2	0,043*"anjing" + 0,038*"orang" + 0,038*"kamu" + 0,031*"banget" + 0,017*"sama" + 0,012*"semua" + 0,010*"pria" + 0,009*"punya" + 0,008*"kayak" + 0,008*"sih"	Ekspresi kemarahan ("banget", "sih") dan makian ("anjing") yang ditujukan baik secara personal ("kamu", "orang") maupun sebagai generalisasi kebencian terhadap suatu kelompok, seperti "semua" "pria".
3	0,024*"guys" + 0,022*"tipu" + 0,018*"bantu" + 0,014*"akun" + 0,012*"kena" + 0,009*"hati" + 0,008*"please" + 0,008*"beli" + 0,007*"nomor" + 0,007*"nama"	Peringatan ("guys", "hati"-hati) dan permintaan tolong ("bantu", "please") karena "kena" "tipu" dalam transaksi "jual" beli atau terkait "akun" yang dicuri atau dipalsukan.

Tabel 8 merangkum interpretasi dari setiap topik yang dihasilkan model, sehingga memudahkan pemahaman konteks dan pola kemunculan kata pada masing-masing klaster.

3.6 Pembahasan

Hasil analisis menunjukkan bahwa percakapan dalam "Komunitas MARAH MARAH" di platform X membentuk tiga klaster topik utama: (1) keluhan emosional sehari-hari, (2) ekspresi kemarahan dan makian, serta (3) isu penipuan dan keamanan digital. Ketiga topik ini mencerminkan dinamika sosial yang beragam dan menunjukkan pola komunikasi khas komunitas berbasis ekspresi emosional. Struktur tematik ini sejalan dengan temuan penelitian sebelumnya bahwa emosi negatif, terutama kemarahan, cenderung mendominasi percakapan berbasis keluhan publik di media sosial. Kemarahan merupakan emosi dengan tingkat *contagiousness* yang tinggi, karena memicu respons cepat dan sering diperkuat oleh dinamika platform [29].

Pertama, topik keluhan emosional sehari-hari menjadi tema yang paling dominan dengan persentase 50,5% dari total percakapan, menggambarkan bagaimana pengguna menyalurkan rasa frustrasi terhadap kehidupan personal, pekerjaan, atau situasi yang berulang. Kemunculan kata seperti *"mau"*, *"banget"*, dan *"marah"* menunjukkan bahwa komunitas ini juga digunakan sebagai ruang *venting* yang bersifat reflektif dan emosional. Temuan ini sejalan dengan studi tentang *venting behavior* di media sosial yang menyatakan bahwa pengguna memanfaatkan platform digital sebagai tempat pelarian untuk mengekspresikan tekanan psikologis dan frustrasi pribadi [30].

Kedua, topik ekspresi kemarahan dan makian menempati proporsi terbesar kedua yakni 35,8%, menonjolkan penggunaan kata bernada agresif seperti *"anjing"*, *"kamu"*, dan *"orang"*. Pola ini mengindikasikan praktik *directed anger* atau *toxic speech* [31], yaitu bentuk kemarahan yang diarahkan kepada individu maupun kelompok tertentu. Konsistensi kata-kata bernada ofensif mengindikasikan bahwa komunitas ini menjadi wadah ekspresi emosional intens yang tidak tersaring oleh norma kesopanan. Hal ini menunjukkan bahwa komunitas tersebut bukan hanya ruang untuk mengeluh, tetapi juga tempat mempertegas identitas emosional kolektif.

Ketiga, topik mengenai penipuan dan keamanan digital dengan proporsi sebesar 13,7%, menunjukkan sisi lain dari komunitas ini, yakni fungsi sebagai ruang peringatan dan solidaritas. Kemunculan kata *"tipu"*, *"akun"*, *"bantu"*, dan *"please"* mengindikasikan bahwa anggota komunitas saling memperingatkan mengenai kasus penipuan, pencurian akun, atau transaksi daring yang berisiko. Temuan ini sejalan dengan laporan penelitian keamanan siber bahwa pengguna media sosial semakin sering mengalami penipuan berbasis akun palsu dan *phishing* [32]. Pola ini memperlihatkan bahwa komunitas tersebut juga berfungsi sebagai ruang berbagi pengalaman negatif terkait transaksi *online* atau ancaman digital.

Berdasarkan *coherence score*, model LDA dengan tiga topik menghasilkan struktur tematik yang paling stabil dan mudah diinterpretasikan. Ketiga topik tersebut memperlihatkan pemisahan semantik yang jelas berdasarkan visualisasi pyLDAvis, sehingga masing-masing dapat diinterpretasikan sebagai klaster tematik yang stabil. Hal ini menunjukkan bahwa meskipun komunitas tersebut homogen dalam konteks "marah-marah," sumber kemarahannya berasal dari konteks yang sangat beragam: mulai dari pengalaman personal, konflik sosial, hingga ancaman digital. Secara keseluruhan, ketiga topik tersebut mencerminkan karakteristik sosial-emosional komunitas *online*: gabungan antara ekspresi personal, respons emosional terhadap situasi eksternal, dan pengalaman menghadapi risiko digital. Temuan ini juga mendukung literatur yang menyatakan bahwa LDA efektif dalam memetakan kelompok emosi atau isu yang sering muncul dalam percakapan publik di media sosial [33].

Selain itu, distribusi kata dalam setiap topik memperlihatkan bahwa komunitas ini tidak hanya berfungsi sebagai sarana meluapkan kemarahan, tetapi juga sebagai mekanisme *sense-making collective*, yaitu proses kolektif dalam memahami dan merespons situasi yang menimbulkan frustrasi. Hal ini memperkuat literatur mengenai dinamika emosi kolektif di media sosial, di mana emosi negatif dapat menyebar secara cepat dan membentuk pola diskursif yang berulang.

Dari perspektif sosiologi digital, keberadaan "Komunitas MARAH MARAH" dapat dilihat sebagai salah satu bentuk *coping mechanism* sosial, tetapi sekaligus membuka peluang terbentuknya polarisasi dan eskalasi konflik verbal. Penggunaan bahasa makian yang intens dapat memperkuat budaya agresivitas, sementara diskusi tentang penipuan menunjukkan adanya kebutuhan edukasi *digital literacy*. Secara keseluruhan, hasil ini menggambarkan ekosistem emosi digital yang kompleks, di mana kemarahan, keluhan, dan solidaritas berjalan beriringan dan membentuk identitas kolektif komunitas tersebut.

Dengan demikian, temuan penelitian ini tidak hanya memberikan gambaran tentang dinamika percakapan dalam komunitas "MARAH MARAH", tetapi juga memperlihatkan bagaimana masyarakat Indonesia menggunakan media sosial sebagai ruang untuk mengelola emosi, mengungkapkan frustrasi, dan memperingatkan orang lain terhadap ancaman digital. Hasil ini memiliki implikasi bagi pemahaman perilaku digital masyarakat, desain kebijakan moderasi platform, serta studi lanjutan mengenai kesehatan mental dan keamanan siber dalam konteks media sosial.

4. KESIMPULAN

Penelitian ini memetakan struktur tematik dari 38.956 *tweet* unik dalam "Komunitas MARAH MARAH" di platform X menggunakan metode *Latent Dirichlet Allocation* (LDA). Melalui serangkaian tahapan *preprocessing* (*remove punctuation*, *case folding*, *tokenizing*, normalisasi, *stopwords removal*, dan *stemming*), data teks disiapkan untuk proses pemodelan. Pengujian sejumlah konfigurasi LDA menunjukkan bahwa tiga topik merupakan jumlah optimal, ditunjukkan oleh nilai *coherence score* tertinggi sebesar 0,5367. Ketiga topik tersebut memiliki distribusi dominasi yang relatif konsisten, yaitu: Topik 1 (keluhan emosional dan curahan personal) sebesar 50,5%, Topik 2 (ekspresi kemarahan dan makian) sebesar 35,8%, dan Topik 3 (isu penipuan serta keamanan digital) sebesar 13,7%. Distribusi ini memberikan gambaran yang lebih jelas tentang bobot relatif masing-masing tema dalam percakapan komunitas, sekaligus memperkuat interpretasi bahwa ruang tersebut digunakan untuk mengekspresikan frustrasi sehari-hari, mengarahkan kemarahan kepada individu atau kelompok tertentu, serta saling memperingatkan mengenai ancaman digital. Secara keseluruhan, hasil penelitian menegaskan bahwa platform media sosial berfungsi tidak hanya sebagai medium komunikasi, tetapi juga sebagai ruang emosional kolektif di mana pengguna mengelola keluhan, mengekspresikan kemarahan, dan membangun solidaritas berbasis pengalaman negatif. Keterbatasan penelitian meliputi belum diterapkannya deteksi bahasa untuk memisahkan *tweet* bilingual, belum dilakukannya analisis temporal untuk mengamati perubahan struktur topik dari waktu ke waktu, serta belum dibandingkannya TF-IDF dengan representasi fitur modern seperti Word2Vec, FastText, atau *embedding* berbasis transformer. Oleh karena itu, penelitian selanjutnya dapat mengintegrasikan *language detection*, analisis longitudinal, serta evaluasi komparatif antar-metode representasi fitur untuk menghasilkan pemetaan topik yang lebih komprehensif, stabil, dan akurat.

REFERENCES

[1] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment Analysis of Twitter Data," *Applied Sciences*, vol. 12, no. 22, Nov. 2022, doi: 10.3390/app122211775.

[2] H. Cam, A. V. Cam, U. Demirel, and S. Ahmed, "Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers," *Heliyon*, vol. 10, no. 1, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23784.

[3] O. Sukmana *et al.*, *Sosiologi Digital: Transformasi Sosial di Era Teknologi*. Yogyakarta: PT Star Digital Publishing, 2025. Accessed: Nov. 19, 2025.

[4] H. R. Kuncoro, K. Hasanah, D. L. Sari, and E. Kurniawati, "Algorithmic Influence: Twitter's Role in Shaping Public Discourse and Amplifying Radical Content," in *Proceedings of the 2nd International Conference on Advance Research in Social and Economic Science (ICARSE 2023)*, Atlantis Press, 2024, pp. 549–553. doi: 10.2991/978-2-38476-247-7_55.

[5] J. Han, S. E. Lee, and M. Cha, "The secret to successful evocative messages: Anger takes the lead in information sharing over anxiety," *Commun Monogr*, vol. 90, no. 4, pp. 545–565, Oct. 2023, doi: 10.1080/03637751.2023.2236183.

[6] I. Z. Al Fatih, R. A. Putera, and Z. H. Umar, "Peran Algoritma Media Sosial dalam Penyebaran Propaganda Politik Digital Menjelang Pemilu," *Jurnal Kajian Stratejik Ketahanan Nasional*, vol. 7, no. 1, Jun. 2024, doi: 10.7454/jkskn.v7i1.10090.

[7] V. Morini, M. Sansoni, G. Rossetti, D. Pedreschi, and C. Castillo, "Participant behavior and community response in online mental health communities: Insights from Reddit," *Comput Human Behav*, vol. 165, p. 108544, Apr. 2025, doi: 10.1016/j.chb.2024.108544.

[8] S. Azzahrani, T. Lukmantoro, and S. R. Manalu, "EKSPRESI EMOSI NEGATIF DALAM MEDIA SOSIAL (STUDI PADA KOMUNITAS 'MARAH-MARAH' DI TWITTER)," *Interaksi Online*, vol. 12, no. 4, pp. 1016–1032, 2024, Accessed: Nov. 19, 2025. [Online]. Available: https://ejournal3.undip.ac.id/index.php/interaksi-online/article/view/47483

[9] I. Z. Al Fatih, "Peran Media Sosial dalam Kampanye Politik di Indonesia Lima Tahun Terakhir: Antara Demokrasi dan Manipulasi Informasi," *COMSERVA : Jurnal Penelitian dan Pengabdian Masyarakat*, vol. 4, no. 7, pp. 2227–2237, Nov. 2024, doi: 10.59141/comserva.v4i7.2611.

[10] Q. Zhang *et al.*, "Analysis of the evolving factors of social media users' emotions and behaviors: a longitudinal study from China's COVID-19 opening policy period," *BMC Public Health*, vol. 23, no. 1, Nov. 2023, doi: 10.1186/s12889-023-17160-y.

[11] M. L. Joshi and N. Kanoongo, "Depression detection using emotional artificial intelligence and machine learning: A closer review," *Mater Today Proc*, vol. 58, no. 1, pp. 217–226, 2022, doi: 10.1016/j.matpr.2022.01.467.

[12] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1–14. doi: 10.18653/v1/S17-2001.

[13] A. Hu and S. Flaxman, "Multimodal Sentiment Analysis To Explore the Structure of Emotions," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2018, pp. 350–358. doi: 10.1145/3219819.3219853.
[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.
[15] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed Tools Appl*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.
[16] B. Jiang, M. Karami, L. Cheng, T. Black, and H. Liu, "Mechanisms and Attributes of Echo Chambers in Social Media," in *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, 2021. doi: 10.48550/arXiv.2106.05401.
[17] G. A. van Kleef and A. H. Fischer, "Emotional collectives: How groups shape emotions and emotions shape groups," *Cogn Emot*, vol. 30, no. 1, pp. 3–19, Jan. 2016, doi: 10.1080/02699931.2015.1081349.
[18] A. M. Raza, N. Aulia, and S. Sopian, "The Echo Chamber Effect: Analisis Sosiologis Peran Algoritma Media Sosial dalam Pembentukan Solidaritas dan Polarisasi Kelompok di Indonesia," *Jurnal Bincang Komunikasi*, vol. 3, no. 1, pp. 39–52, 2025, Accessed: Nov. 19, 2025. [Online]. Available: https://jurnal.umj.ac.id/index.php/JBK/article/view/27736
[19] Mahyuddin, *Sosiologi Komunikasi: Dinamika Relasi Sosial di dalam Era Virtualitas*. Makassar: Shofia - CV Loe, 2019.
[20] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-76917-0.
[21] D. Rifaldi, A. Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, Apr. 2023, doi: 10.51454/decode.v3i2.131.
[22] A. R. Lubis, Y. Y. Lase, D. A. Rahman, and D. Witarsyah, "Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models," *Ingénierie des systèmes d information*, vol. 28, no. 5, pp. 1335–1342, Oct. 2023, doi: 10.18280/isi.280522.
[23] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ Res Methods*, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.
[24] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, and J. Teixeira, "Analysis of Document Pre-Processing Effects in Text and Opinion Mining," *Information*, vol. 9, no. 4, p. 100, Apr. 2018, doi: 10.3390/info9040100.
[25] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, 2018.
[26] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
[27] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front Artif Intell*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.
[28] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," *Neurocomputing*, vol. 72, no. 7–9, pp. 1775–1781, Mar. 2009, doi: 10.1016/j.neucom.2008.06.011.
[29] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7313–7318, Jul. 2017, doi: 10.1073/pnas.1618923114.
[30] S. Steinert and M. J. Dennis, "Emotions and Digital Well-Being: on Social Media's Emotional Affordances," *Philos Technol*, vol. 35, pp. 1–21, Jun. 2022, doi: 10.1007/s13347-022-00530-6.
[31] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," in *1st Workshop on Abusive Language Online*, May 2017. Accessed: Nov. 22, 2025. [Online]. Available: https://arxiv.org/abs/1705.09899
[32] A. Hamid, M. Alam, H. Sheherin, and A.-S. K. Pathan, "Cyber Security Concerns in Social Networking Service," *International Journal of Communication Networks and Information Security*, vol. 12, no. 2, pp. 198–212, 2020, Accessed: Nov. 22, 2025. [Online]. Available: https://www.proquest.com/docview/2440678475?fromopenview=true&pq-origsite=gscholar&sourcetype=Scholarly%20Journals
[33] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed Tools Appl*, vol. 78, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.