Journal of Informatics and Computer Science Vol. 12, No. 1 April 2026
Universitas Ubudiyah Indonesia
e-ISSN : 2615-5346

ANALISIS TOPIK KELUHAN MASYARAKAT TERHADAP PENANGANAN BANJIR DI MEDIA SOSIAL MENGGUNAKAN BERtopic

TOPIC ANALYSIS OF PUBLIC COMPLAINTS REGARDING FLOOD MANAGEMENT ON SOCIAL MEDIA USING BERtopic

Desita Ria Yusian TB[1], Mazidul Ihsan[2], Mahendar Dwi Payana[3], M.Bayu Wibawa[4]
[1,2,3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Ubudiyah Indonesia,
[4]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Ubudiyah Indonesia,
Jl. Alue Naga, Tibang. Kec. Syiah Kuala, Banda Aceh, Indonesia
Coresponding Author :desita@uui.ac.id

***Abstrak*—** Banjir merupakan salah satu bencana yang menimbulkan berbagai permasalahan dalam penanganannya, yang kemudian mendorong masyarakat menyampaikan keluhan, kritik, dan tanggapan melalui media sosial. Data komentar pada media sosial memiliki jumlah yang besar, bersifat tidak terstruktur, dan menggunakan bahasa yang beragam sehingga sulit dianalisis secara manual. Penelitian ini bertujuan mengidentifikasi dan memetakan topik-topik keluhan masyarakat terhadap penanganan banjir melalui media sosial YouTube dan TikTok menggunakan metode BERTopic. Data dikumpulkan dari enam sumber konten pada kedua platform dan diproses melalui tahapan preprocessing yang meliputi case folding, cleaning, normalization, tokenization, dan stopword removal. Selanjutnya, pemodelan topik dilakukan menggunakan Sentence-BERT dengan model *paraphrase-multilingual-MiniLM-L12-v2*, UMAP, HDBSCAN, dan c-TF-IDF. Dari 18.447 komentar hasil preprocessing, BERTopic menghasilkan 268 topik yang kemudian diinterpretasikan dan dipilih 10 topik yang paling relevan, yaitu Sebaran Banjir, Kebutuhan Air Bersih, Permintaan Bantuan, Relawan, Masyarakat Terdampak, Akses Bantuan, Penyaluran Bantuan, Kecepatan Penanganan, Status Bencana, dan Kritik Pemerintah. Evaluasi menggunakan *Normalized Pointwise Mutual Information* (NPMI) menghasilkan nilai 0,1962. Hasil penelitian menunjukkan bahwa BERTopic dapat digunakan untuk memetakan berbagai isu keluhan masyarakat secara lebih terstruktur, dengan isu bantuan, kondisi bencana, kebutuhan air bersih, relawan, dan kritik terhadap penanganan pemerintah menjadi bagian penting dalam pembentukan topik.

Kata kunci: ***BERTopic, keluhan masyarakat, banjir, media sosial, topic modeling***

***Abstract*—** *Flooding is one of the disasters that causes various challenges in disaster management, leading people to express complaints, criticism, and responses through social media. Social media comments are large in volume, unstructured, and characterized by diverse language, making them difficult to analyze manually. This study aims to identify and map public complaint topics regarding flood management on YouTube and TikTok using the BERTopic method. Data were collected from six content sources across the two platforms and processed through preprocessing stages consisting of case folding, cleaning, normalization, tokenization, and stopword removal. Topic modeling was then performed using Sentence-BERT with the paraphrase-multilingual-MiniLM-L12-v2 model, UMAP, HDBSCAN, and c-TF-IDF. From 18,447 preprocessed comments, BERTopic generated 268 topics, which were subsequently interpreted and reduced to 10 topics most relevant to the research context: Flood Distribution, Clean Water Needs, Requests for Assistance, Volunteers, Affected Communities, Access to Assistance, Aid Distribution, Response Speed, Disaster Status, and Government Criticism. Model evaluation using Normalized Pointwise Mutual Information (NPMI) produced an overall score of 0.1962. The results indicate that BERTopic can be used to systematically map various public complaint issues, with assistance, disaster conditions, clean water needs, volunteer involvement, and criticism of government response emerging as important themes in the identified topics.*

Keywords: ***BERTopic, public complaints, flooding, social media, topic modeling***

I. PENDAHULUAN

Bencana banjir merupakan salah satu peristiwa yang dapat menimbulkan dampak luas terhadap kehidupan masyarakat, baik dalam aspek sosial, ekonomi, infrastruktur, maupun aktivitas sehari-hari. Dalam kondisi tersebut, proses penanganan bencana membutuhkan informasi yang cepat dan relevan mengenai kondisi lapangan, kebutuhan masyarakat, serta berbagai permasalahan yang muncul selama tanggap darurat. Ketersediaan informasi dari masyarakat menjadi penting karena informasi yang berasal langsung dari individu yang terdampak dapat memberikan gambaran mengenai kondisi aktual yang tidak selalu tersedia melalui sumber informasi formal. Perkembangan teknologi digital kemudian memungkinkan masyarakat untuk menyampaikan informasi, pengalaman, keluhan, dan kritik secara langsung melalui media sosial. Pemanfaatan data yang berasal dari masyarakat melalui media sosial bahkan telah dipandang sebagai salah satu sumber informasi yang potensial untuk mendukung respons bencana berbasis data [1].

Media sosial tidak hanya berfungsi sebagai sarana komunikasi dan penyebaran informasi, tetapi juga berkembang menjadi ruang partisipasi publik dalam situasi bencana. Melalui fitur komentar, masyarakat dapat menyampaikan pengalaman, kondisi lingkungan, kebutuhan mendesak, maupun penilaian terhadap proses penanganan

bencana. Penelitian mengenai komunikasi bencana di Indonesia menunjukkan bahwa media sosial telah digunakan dalam komunikasi risiko bencana, meskipun masih terdapat tantangan berupa kredibilitas informasi, misinformasi, dan kesenjangan digital [2]. Dalam konteks banjir, data media sosial juga dapat memberikan informasi yang relatif cepat mengenai lokasi dan dampak banjir serta respons masyarakat terhadap kondisi yang terjadi [3], [4]. Penelitian mengenai respons masyarakat terhadap banjir di Jakarta menunjukkan bahwa percakapan masyarakat di media sosial dapat memberikan gambaran mengenai pengalaman dan perhatian publik terhadap pengelolaan banjir [5].

YouTube dan TikTok menjadi dua *platform* yang memiliki karakteristik relevan untuk mengamati respons publik terhadap peristiwa banjir. Kedua *platform* memungkinkan pengguna memberikan komentar secara langsung terhadap konten video yang berkaitan dengan kondisi bencana. Komentar tersebut dapat berisi informasi mengenai kondisi wilayah terdampak, permintaan bantuan, kebutuhan dasar, kritik terhadap penanganan, maupun apresiasi terhadap pihak yang terlibat dalam penanggulangan bencana. Dalam penelitian ini, sumber data terdiri atas enam konten video pada YouTube dan TikTok yang berkaitan dengan peristiwa banjir di Aceh. Data komentar dari sumber-sumber tersebut selanjutnya disusun menjadi dataset terstruktur yang memuat atribut *author*, *comment*, *likes*, *replies*, *published_at*, *platform*, dan *source_video*. Dengan demikian, komentar masyarakat tidak hanya dipandang sebagai opini, tetapi juga sebagai data tekstual yang dapat dianalisis untuk mengidentifikasi pola permasalahan yang berkembang dalam percakapan publik.

Meskipun memiliki nilai informasi yang tinggi, komentar media sosial memiliki karakteristik yang kompleks. Data umumnya berjumlah besar, tidak terstruktur, bersifat pendek, menggunakan bahasa informal, serta mengandung variasi penulisan, singkatan, dan ekspresi subjektif. Kondisi tersebut menyebabkan identifikasi isu secara manual membutuhkan waktu dan sulit dilakukan secara konsisten. Permasalahan serupa juga ditemukan dalam penelitian analisis media sosial untuk kebencanaan, karena besarnya volume data membuat proses penyaringan informasi yang relevan menjadi tantangan tersendiri [6]. Oleh karena itu, diperlukan pendekatan pengolahan teks yang mampu mengorganisasi data komentar ke dalam kelompok isu berdasarkan kesamaan makna, sehingga informasi yang tersebar dapat dipahami secara lebih sistematis.

Salah satu pendekatan yang dapat digunakan untuk mengatasi permasalahan tersebut adalah *Natural Language Processing (NLP)* melalui pemodelan topik. Pemodelan topik bertujuan menemukan struktur atau tema laten yang terdapat dalam sekumpulan dokumen tanpa harus menentukan kategori secara manual terlebih dahulu. Salah satu metode yang berkembang untuk tujuan tersebut adalah BERTopic, yang memanfaatkan representasi embedding berbasis *transformer* untuk menangkap hubungan semantik antardokumen. BERTopic melakukan pembentukan representasi dokumen, pengelompokan embedding, dan pembentukan representasi topik menggunakan *class-based Term Frequency–Inverse Document Frequency* (c-TF-IDF) [7]. Pendekatan tersebut memungkinkan teks dikelompokkan berdasarkan kemiripan semantik sehingga lebih sesuai untuk data media sosial yang pendek dan tidak terstruktur.

Pada implementasinya, penelitian ini menggunakan *Sentence-BERT* (SBERT) untuk menghasilkan representasi semantik komentar. SBERT dikembangkan untuk menghasilkan *sentence embedding* yang dapat dibandingkan menggunakan *cosine similarity*, sehingga sesuai untuk kebutuhan pencarian kemiripan dan pengelompokan teks [8]. Embedding yang dihasilkan selanjutnya direduksi dimensinya menggunakan *Uniform Manifold Approximation and Projection* (UMAP) sebelum dilakukan pengelompokan menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). UMAP merupakan metode reduksi dimensi nonlinier yang dirancang untuk mempertahankan struktur penting dari data berdimensi tinggi [9], sedangkan HDBSCAN merupakan pendekatan *density-based clustering* yang mampu membentuk hierarki cluster berdasarkan kepadatan data [10]. Kombinasi tersebut memungkinkan BERTopic mengelompokkan komentar berdasarkan kedekatan semantik dan kemudian menghasilkan kata-kata representatif untuk setiap topik menggunakan c-TF-IDF.

Sejumlah penelitian sebelumnya telah menunjukkan potensi analisis media sosial untuk memahami respons masyarakat terhadap bencana. Penelitian mengenai banjir Jakarta, misalnya, telah menggunakan data Twitter untuk mengidentifikasi respons masyarakat dan membangun taksonomi respons terhadap bencana banjir [5]. Penelitian lain mengenai bencana banjir juga menggunakan pemodelan topik berbasis LDA untuk mengidentifikasi tema-tema utama dalam percakapan publik [11]. Di sisi lain, BERTopic telah digunakan pada berbagai permasalahan analisis teks karena kemampuannya menghasilkan representasi topik berbasis konteks dan hubungan semantik [7]. Namun demikian, berdasarkan kajian pustaka yang digunakan dalam penelitian ini, penerapan BERTopic secara khusus untuk memetakan keluhan masyarakat terhadap penanganan banjir melalui komentar YouTube dan TikTok, khususnya dalam konteks bencana banjir di Aceh**,** masih terbatas. Penelitian terdahulu lebih banyak berfokus pada analisis sentimen, komunikasi bencana, Twitter, atau pemodelan topik menggunakan pendekatan konvensional. Dengan demikian, terdapat celah penelitian berupa kebutuhan untuk memanfaatkan pemodelan topik berbasis embedding yang mampu menangkap hubungan semantik dari komentar masyarakat yang pendek, informal, dan tidak terstruktur.

Berdasarkan kesenjangan tersebut, penelitian ini berfokus pada pemodelan topik keluhan masyarakat terhadap penanganan banjir pada media sosial YouTube dan TikTok menggunakan BERTopic**.** Data komentar dikumpulkan melalui *web scraping* menggunakan YouTube Data API v3 dan Apify TikTok Comments Scraper, kemudian melalui tahapan *preprocessing* sebelum dilakukan pemodelan topik. Kualitas topik yang dihasilkan dievaluasi menggunakan Normalized Pointwise Mutual Information (NPMI) sebagai ukuran *topic coherence*. NPMI digunakan untuk mengukur keterkaitan semantik antarkata dalam suatu topik, dengan nilai yang lebih tinggi menunjukkan keterkaitan kata yang lebih kuat [12].

Penelitian ini bertujuan untuk (1) mengumpulkan dan mengolah data komentar masyarakat terkait penanganan banjir dari media sosial YouTube dan TikTok, (2) memodelkan topik keluhan masyarakat menggunakan metode BERTopic, dan (3) mengidentifikasi serta mendeskripsikan topik-topik utama yang terbentuk sehingga pola keluhan masyarakat terhadap penanganan banjir dapat dipahami secara lebih sistematis**.** Hasil penelitian diharapkan dapat memberikan pemetaan isu berbasis data mengenai keluhan masyarakat dan menjadi informasi pendukung bagi penelitian maupun pengembangan analisis opini publik dan komunikasi kebencanaan berbasis *Natural Language Processing*.

II. STUDI PUSTAKA

2.1 Natural Language Processing

Natural Language Processing (NLP) merupakan bidang dalam kecerdasan buatan yang berfokus pada pemrosesan dan pemahaman bahasa manusia oleh komputer. NLP memungkinkan data teks yang bersifat tidak terstruktur diolah menjadi informasi yang dapat dianalisis secara komputasional. Dalam perkembangannya, NLP digunakan pada berbagai tugas seperti klasifikasi teks, analisis sentimen, pencarian informasi, *semantic similarity*, dan pemodelan topik [13].

Data media sosial memiliki karakteristik yang berbeda dengan dokumen formal karena umumnya berupa teks pendek, tidak terstruktur, menggunakan bahasa informal, singkatan, variasi ejaan, serta ekspresi subjektif. Karakteristik tersebut menjadi tantangan dalam proses pengolahan teks dan membutuhkan tahapan *preprocessing* agar data lebih konsisten sebelum digunakan dalam pemodelan. Pada penelitian ini, NLP digunakan sebagai pendekatan untuk mengolah komentar masyarakat dari YouTube dan TikTok melalui tahapan *case folding*, *cleaning*, *normalization*, *tokenization*, dan *stopword removal*.

Selain *preprocessing*, NLP modern memanfaatkan representasi teks berbasis *embedding* untuk menangkap hubungan semantik. Salah satu pendekatan yang digunakan adalah *Sentence-BERT* (SBERT). SBERT merupakan pengembangan BERT yang menggunakan arsitektur *siamese network* untuk menghasilkan representasi kalimat dalam bentuk vektor sehingga kemiripan semantik antarkalimat dapat dihitung secara efisien menggunakan *cosine similarity* [14]. Dalam penelitian ini, SBERT digunakan sebagai tahap awal pembentukan representasi semantik komentar sebelum dilakukan pemodelan topik. Model yang digunakan adalah *paraphrase-multilingual-MiniLM-L12-v2* dengan dimensi embedding 384.

2.2 Topic Modeling

Topic modeling merupakan pendekatan *unsupervised learning* yang digunakan untuk menemukan struktur topik atau tema laten dari sekumpulan dokumen. Tujuan utamanya adalah mengelompokkan dokumen berdasarkan pola dan hubungan kata yang muncul sehingga kumpulan teks yang besar dapat direpresentasikan dalam sejumlah topik yang lebih mudah dipahami. Topic modeling banyak digunakan untuk mengeksplorasi koleksi dokumen tanpa memerlukan pemberian label secara manual sebelumnya.
Dalam perkembangannya, topic modeling tidak hanya menggunakan pendekatan berbasis frekuensi kata, tetapi juga berkembang menuju pendekatan berbasis representasi semantik. Hal ini menjadi penting ketika diterapkan pada data media sosial yang pendek dan memiliki variasi bahasa tinggi. Penelitian mengenai topic modeling pada teks pendek menunjukkan bahwa hubungan semantik dan konteks menjadi aspek penting karena metode konvensional yang mengandalkan *word co-occurrence* menghadapi tantangan pada dokumen dengan jumlah kata yang terbatas [15]. Secara umum, kualitas topic modeling tidak hanya ditentukan oleh jumlah topik yang dihasilkan, tetapi juga oleh keterinterpretasian topik, kemampuan mengelompokkan dokumen secara bermakna, serta keterkaitan semantik kata-kata dalam setiap topik. Oleh karena itu, evaluasi *topic coherence* diperlukan untuk mengetahui sejauh mana kata-kata yang membentuk suatu topik memiliki keterkaitan yang konsisten. Penelitian ini menggunakan Normalized Pointwise Mutual Information (NPMI) sebagai ukuran *topic coherence* untuk mengevaluasi kualitas topik yang dihasilkan.

2.3 BERTopic

BERTopic merupakan metode *neural topic modeling* yang memanfaatkan representasi *embedding* dari model bahasa berbasis Transformer untuk mengelompokkan dokumen berdasarkan kemiripan semantik. Berbeda dengan pendekatan topic modeling konvensional yang sangat bergantung pada frekuensi kemunculan kata, BERTopic membentuk representasi dokumen dalam ruang vektor, melakukan *clustering* terhadap representasi tersebut, kemudian menghasilkan representasi topik menggunakan *class-based Term Frequency–Inverse Document Frequency* (c-TF-IDF) [16].

Secara umum, proses BERTopic terdiri atas empat komponen utama, yaitu document embedding, dimensionality reduction, clustering, dan topic representation. Tahap pertama mengubah setiap dokumen menjadi representasi vektor menggunakan model *sentence embedding*. Selanjutnya, embedding direduksi dimensinya menggunakan *Uniform Manifold Approximation and Projection (UMAP)* untuk mempertahankan struktur penting dari data berdimensi tinggi. Setelah itu, *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) digunakan untuk mengelompokkan dokumen berdasarkan kepadatan dan kemiripan representasi. Tahap terakhir menggunakan c-TF-IDF untuk menghasilkan kata atau frasa yang paling representatif pada setiap topik [16]. Pada penelitian ini, tahap *embedding* menggunakan *Sentence-BERT* dengan model *paraphrase-multilingual-MiniLM-L12-v2*. Embedding tersebut kemudian diproses menggunakan UMAP dan HDBSCAN, sedangkan representasi topik dibentuk menggunakan c-TF-IDF. Rangkaian tersebut memungkinkan komentar masyarakat dikelompokkan berdasarkan kesamaan makna, bukan hanya berdasarkan kemunculan kata yang sama. Keunggulan lain BERTopic adalah kemampuannya menghasilkan representasi topik yang relatif kontekstual dan mudah diinterpretasikan. Hal tersebut relevan dengan penelitian ini karena komentar masyarakat terkait banjir bersifat pendek, informal, dan tidak terstruktur. Dengan pendekatan berbasis *embedding*, komentar yang memiliki konteks semantik serupa dapat ditempatkan pada kelompok topik yang sama meskipun menggunakan kata yang berbeda.

2.4 Penelitian Terdahulu

Beberapa penelitian sebelumnya menunjukkan penggunaan BERTopic maupun metode analisis teks untuk memahami opini dan respons masyarakat. Penelitian Wahyuni et al. (2025) berjudul *"Analisis Persepsi Masyarakat terhadap Program Makan Bergizi Gratis Menggunakan BERTopic pada Komentar YouTube"* menerapkan BERTopic, *Sentence-BERT*, UMAP, dan HDBSCAN terhadap 19.843 komentar YouTube dan menghasilkan 10 topik utama yang merepresentasikan persepsi masyarakat terhadap program tersebut. Penelitian ini menunjukkan bahwa BERTopic dapat digunakan untuk mengidentifikasi struktur topik dari komentar masyarakat dalam jumlah besar. Namun, objek penelitian masih berfokus pada persepsi masyarakat terhadap suatu program pemerintah dan belum secara khusus mengidentifikasi keluhan pada konteks kebencanaan.

Penelitian Aryani et al. (2024) berjudul *"Identifikasi Topik Dominan pada Pemberitaan Pemilu 2024 Menggunakan BERTopic"* menggunakan pendekatan BERTopic berbasis *embedding* untuk mengidentifikasi topik dominan pada pemberitaan Pemilu 2024. Hasil penelitian menunjukkan bahwa BERTopic mampu menghasilkan representasi topik yang lebih kontekstual karena mempertimbangkan hubungan semantik antarkata. Meskipun demikian, data yang digunakan berupa artikel berita yang relatif lebih formal dan terstruktur dibandingkan komentar media sosial. Oleh karena itu, karakteristik data penelitian tersebut berbeda dengan komentar YouTube dan TikTok yang cenderung pendek, informal, dan memiliki variasi ekspresi yang tinggi.

Sementara itu, Erwanda et al. (2025) melakukan penelitian berjudul *"Analisis Respons Masyarakat terhadap Bencana Banjir di Aceh Menggunakan Pendekatan Lexicon-Based Sentiment Analysis"*. Penelitian tersebut menggunakan analisis sentimen berbasis *lexicon* untuk mengidentifikasi respons masyarakat terhadap bencana banjir di Aceh. Hasilnya menunjukkan bahwa opini masyarakat di media sosial tidak hanya mengandung ekspresi emosi, tetapi juga kritik, harapan, dan informasi mengenai kondisi lapangan. Namun, pendekatan analisis sentimen terutama menghasilkan informasi mengenai polaritas opini, seperti positif, negatif, dan netral, sehingga belum secara langsung menunjukkan struktur tema atau isu spesifik yang menjadi dasar keluhan masyarakat.

Berdasarkan penelitian terdahulu tersebut, terdapat perbedaan pada objek, karakteristik data, dan pendekatan analisis. Penelitian Wahyuni et al. menunjukkan kemampuan BERTopic dalam menganalisis komentar YouTube, sedangkan Aryani et al. menunjukkan kemampuan BERTopic dalam mengidentifikasi topik pada teks berita. Di sisi lain, penelitian Erwanda et al. telah mengkaji respons masyarakat terhadap banjir di Aceh, tetapi menggunakan analisis sentimen yang belum secara khusus memetakan struktur topik keluhan. Dengan demikian, posisi penelitian ini adalah mengintegrasikan pendekatan BERTopic dengan data komentar YouTube dan TikTok untuk mengidentifikasi serta memetakan topik keluhan masyarakat terhadap penanganan banjir. Pendekatan tersebut diarahkan untuk menghasilkan peta isu yang lebih spesifik dibandingkan sekadar mengetahui polaritas sentimen masyarakat. Hal ini sejalan dengan tujuan penelitian dalam skripsi, yaitu mengidentifikasi dan mendeskripsikan topik-topik dominan sehingga pola keluhan masyarakat terhadap penanganan banjir dapat dipahami secara lebih sistematis.

III. METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif berbasis komputasi dengan teknik *text mining* dan *Natural Language Processing (NLP)*. Fokus penelitian adalah mengidentifikasi dan memetakan topik keluhan masyarakat terhadap penanganan banjir berdasarkan komentar pada media sosial menggunakan metode BERTopic. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, pembentukan *sentence embedding*

menggunakan *Sentence-BERT* (SBERT), reduksi dimensi menggunakan UMAP, *clustering* menggunakan HDBSCAN, pembentukan representasi topik menggunakan c-TF-IDF, dan evaluasi kualitas topik menggunakan *topic coherence* dengan metrik *Normalized Pointwise Mutual Information (NPMI).*

3.1 Data Penelitian

Data penelitian berupa komentar masyarakat yang diperoleh dari dua *platform* media sosial, yaitu YouTube dan TikTok. Pemilihan sumber data didasarkan pada keterkaitannya dengan peristiwa banjir, jumlah komentar yang relatif banyak, dan aktivitas diskusi yang tinggi. Data tidak hanya berasal dari akun media berita, tetapi juga mencakup konten dari kreator atau akun lain yang relevan dengan topik penelitian.

Pengumpulan data dilakukan menggunakan teknik *web scraping*. Data YouTube diperoleh menggunakan YouTube Data API v3, sedangkan data TikTok diperoleh menggunakan Apify TikTok Scraper API. Dari proses pengumpulan tersebut diperoleh 8.323 komentar YouTube dan 10.124 komentar TikTok, sehingga jumlah keseluruhan dataset adalah 18.447 komentar. Data kemudian digabungkan dan disimpan dalam format Microsoft Excel untuk proses pengolahan lebih lanjut.

Tabel 1. Sumber dan Jumlah Data

Platform	**Jumlah Komentar**
YouTube	8.323
TikTok	10.124
Total	**18.447**

Dataset disusun dengan beberapa atribut, yaitu *author*, *comment*, *likes*, *replies*, *published_at*, *platform*, dan *source_video*. Atribut *comment* digunakan sebagai data utama dalam proses pemodelan topik, sedangkan atribut lainnya digunakan sebagai informasi pendukung dataset.

3.2 Preprocessing Data

Komentar media sosial memiliki karakteristik berupa teks pendek, tidak terstruktur, penggunaan bahasa informal, variasi ejaan, serta keberadaan simbol dan elemen nonteks. Oleh karena itu, dilakukan *preprocessing* untuk mengurangi *noise* dan menghasilkan teks yang lebih konsisten sebelum masuk ke tahap pembentukan *embedding*. Tahapan *preprocessing* terdiri atas case folding, cleaning, normalization, tokenization, dan stopword removal.

Case folding dilakukan dengan mengubah seluruh huruf menjadi huruf kecil sehingga variasi penggunaan huruf kapital tidak memengaruhi proses analisis. Selanjutnya, cleaning digunakan untuk menghapus elemen yang tidak diperlukan, seperti URL, *mention*, *hashtag*, emoji, angka, tanda baca, karakter khusus, spasi berlebih, serta komentar kosong atau duplikat. Tahap normalization dilakukan untuk menyeragamkan penulisan dengan mengubah kata tidak baku, singkatan, bahasa gaul, dan kesalahan penulisan menjadi bentuk yang lebih baku menggunakan kamus normalisasi bahasa Indonesia. Setelah itu, tokenization digunakan untuk memecah kalimat menjadi token kata. Tahap terakhir adalah stopword removal, yaitu menghilangkan kata-kata umum yang tidak memberikan kontribusi signifikan terhadap pembentukan topik. Setelah seluruh tahapan *preprocessing* dilakukan, jumlah data yang digunakan dalam pemodelan tetap sebanyak 18.447 komentar.

3.3 *Sentence-BERT (SBERT)*

Tahap berikutnya adalah mengubah setiap komentar menjadi representasi numerik berupa *sentence embedding*. Penelitian ini menggunakan *Sentence-BERT (SBERT*) dengan model *paraphrase-multilingual-MiniLM-L12-v2*. Model tersebut dipilih karena mendukung representasi multibahasa dan mampu menangkap kemiripan semantik antarkalimat. Model menghasilkan representasi vektor berdimensi 384 untuk setiap komentar. *Embedding* memungkinkan komentar yang memiliki makna atau konteks serupa memiliki representasi vektor yang berdekatan. Dengan demikian, proses pemodelan tidak hanya bergantung pada kesamaan kata secara literal, tetapi juga mempertimbangkan hubungan semantik antar-komentar. SBERT sendiri merupakan pengembangan BERT yang dirancang untuk menghasilkan *sentence embedding* yang dapat dibandingkan menggunakan *cosine similarity* [1]. Dalam konteks penelitian ini, embedding dari 18.447 komentar menjadi masukan untuk tahap reduksi dimensi dan *clustering* dalam kerangka BERTopic.

3.4 Dimensionality Reduction menggunakan UMAP

Representasi *embedding* yang dihasilkan SBERT memiliki dimensi yang relatif tinggi sehingga perlu direduksi sebelum proses *clustering*. Penelitian ini menggunakan *Uniform Manifold Approximation and Projection* (UMAP) sebagai metode reduksi dimensi. UMAP digunakan untuk menyederhanakan representasi vektor dengan tetap mempertahankan struktur dan hubungan penting antardata. Pada pipeline BERTopic, reduksi dimensi dilakukan sebelum *clustering* agar representasi embedding lebih sesuai untuk proses pengelompokan berbasis kepadatan [2]. Dokumentasi BERTopic juga menempatkan UMAP sebagai tahap reduksi dimensi setelah pembentukan *embedding* dan sebelum HDBSCAN. Dalam penelitian ini, hasil reduksi dimensi UMAP selanjutnya digunakan sebagai masukan untuk algoritma HDBSCAN.

3.5 Clustering menggunakan HDBSCAN

Setelah reduksi dimensi, dilakukan proses *clustering* menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN).*

HDBSCAN merupakan algoritma *density-based clustering* yang mengelompokkan data berdasarkan kepadatan dan mampu mengidentifikasi data yang tidak termasuk dalam kelompok tertentu sebagai *noise* atau *outlier*. Penggunaan HDBSCAN penting dalam penelitian ini karena jumlah topik tidak harus ditentukan terlebih dahulu secara manual. Dokumen yang memiliki kemiripan semantik dan kepadatan yang cukup dapat membentuk suatu kelompok, sedangkan komentar yang tidak memiliki kedekatan yang memadai dapat ditempatkan sebagai *outlier*. Dalam pipeline BERTopic, HDBSCAN digunakan untuk mengelompokkan embedding hasil reduksi dimensi menjadi kelompok dokumen yang selanjutnya diperlakukan sebagai topik. Hasil proses pemodelan pada penelitian ini menghasilkan 268 topik dari 18.447 komentar yang telah melalui *preprocessing*.

3.6 Topic Representation menggunakan c-TF-IDF

Setelah dokumen dikelompokkan oleh HDBSCAN, setiap kelompok perlu direpresentasikan menggunakan kata-kata yang paling relevan. Pada tahap ini digunakan *class-based Term Frequency–Inverse Document Frequency (c-TF-IDF).* Sebelum proses c-TF-IDF, setiap *cluster* direpresentasikan dalam bentuk *Bag of Words* menggunakan *CountVectorizer*. Selanjutnya, seluruh dokumen dalam suatu *cluster* diperlakukan sebagai satu kelompok dokumen sehingga dapat dihitung bobot kata yang paling representatif terhadap kelompok tersebut. c-TF-IDF merupakan komponen penting dalam BERTopic karena digunakan untuk menghasilkan kata kunci yang merepresentasikan setiap topik. Dengan demikian, hasil *clustering* yang sebelumnya berupa kelompok dokumen dapat diinterpretasikan menjadi tema atau isu berdasarkan kata-kata dengan bobot tertinggi [3]. BERTopic secara khusus menggunakan c-TF-IDF untuk membentuk representasi topik dari kelompok dokumen yang telah terbentuk. Berdasarkan hasil c-TF-IDF, seluruh 268 topik kemudian diinterpretasikan berdasarkan kata-kata representatifnya. Selanjutnya dipilih 10 topik yang paling relevan dengan konteks keluhan masyarakat terhadap penanganan banjir. Proses *semantic topic matching* menggunakan SBERT, *cosine similarity*, dan *keyword overlap* digunakan untuk memastikan kesesuaian antara nama topik dan *Topic ID* yang dihasilkan model.

3.7 Evaluasi Topic Coherence

Evaluasi dilakukan untuk mengetahui kualitas topik yang dihasilkan oleh BERTopic. Penelitian ini menggunakan *Normalized Pointwise Mutual Information (NPMI)* sebagai metrik *topic coherence*. NPMI mengukur tingkat keterkaitan atau kemunculan bersama kata-kata representatif dalam suatu topik. Nilai yang semakin tinggi menunjukkan hubungan antarkata yang semakin kuat dan koherensi topik yang semakin baik. Nilai NPMI dihitung untuk setiap pasangan kata kunci dalam suatu topik, kemudian dirata-ratakan untuk memperoleh nilai koherensi per topik. Evaluasi keseluruhan dilakukan terhadap 265 topik, dengan mengecualikan *outlier* atau Topic -1. Hasil evaluasi terhadap 18.447 komentar menghasilkan nilai NPMI sebesar 0,1962.

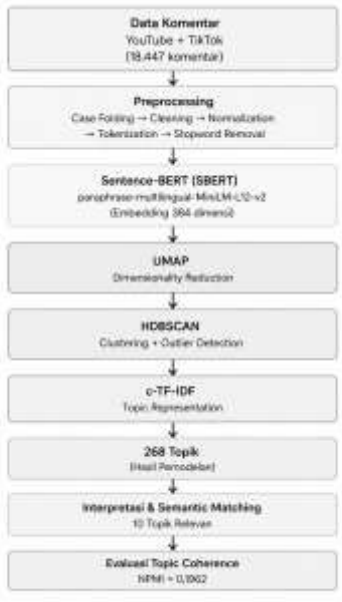
Tabel 2. Evaluasi Topic Coherence

Parameter	Nilai
Jumlah komentar	18.447
Topik hasil BERTopic	268
Topik dievaluasi	265
Metrik	NPMI
Nilai NPMI	0,1962

Nilai NPMI sebesar 0,1962 menunjukkan tingkat koherensi yang moderat/cukup, yang dalam penelitian ini dikaitkan dengan karakteristik komentar media sosial yang pendek, tidak terstruktur, dan menggunakan bahasa informal.

3.8 Alur Metode Penelitian

Alur metode penelitian yang digunakan dalam penelitian ini ditunjukkan pada Gambar 1. Proses penelitian diawali dengan pengumpulan 18.447 komentar dari *platform* YouTube dan TikTok yang berkaitan dengan penanganan banjir.



Gambar 1. Alur Pemodelan Topik Keluhan Masyarakat

Data yang terkumpul kemudian melalui tahap *preprocessing* yang meliputi *case folding, cleaning, normalization, tokenization,* dan *stopword removal* untuk menghasilkan teks yang lebih bersih dan konsisten. Komentar hasil *preprocessing* selanjutnya diubah menjadi representasi numerik menggunakan *Sentence-BERT* (SBERT) dengan model *paraphrase-multilingual-MiniLM-*

L12-v2. Representasi *embedding* yang dihasilkan kemudian direduksi dimensinya menggunakan UMAP dan dikelompokkan menggunakan HDBSCAN berdasarkan kepadatan dan kemiripan data. Setelah proses *clustering*, c-TF-IDF digunakan untuk menghasilkan kata-kata yang paling representatif pada setiap kelompok sehingga dapat diinterpretasikan sebagai topik.

Hasil pemodelan menghasilkan 268 topik dari 18.447 komentar. Seluruh topik kemudian diinterpretasikan berdasarkan kata kunci yang dihasilkan c-TF-IDF, kemudian dilakukan *semantic topic matching* menggunakan SBERT, *cosine similarity*, dan *keyword overlap* untuk memastikan kesesuaian antara nama topik dan *Topic ID*. Dari proses tersebut dipilih 10 topik yang paling relevan dengan konteks keluhan masyarakat terhadap penanganan banjir. Tahap terakhir adalah evaluasi kualitas topik menggunakan *topic coherence* dengan metrik *Normalized Pointwise Mutual Information (NPMI)*. Evaluasi dilakukan terhadap 265 topik setelah *outlier* dikeluarkan dan menghasilkan nilai NPMI sebesar 0,1962. Nilai tersebut digunakan untuk menggambarkan tingkat keterkaitan kata-kata dalam topik yang dihasilkan oleh model BERTopic.

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Pengumpulan Data

Data penelitian diperoleh dari komentar pada *platform* YouTube dan TikTok yang berkaitan dengan peristiwa dan penanganan banjir. Pengumpulan data dilakukan menggunakan *YouTube Data API v3* dan *Apify TikTok Scraper API*. Data yang terkumpul kemudian disusun dalam dataset terstruktur dan digunakan sebagai masukan untuk proses *preprocessing* dan pemodelan topik. Setelah proses pengumpulan dan penyiapan data, diperoleh 18.447 komentar yang digunakan dalam proses pemodelan, terdiri atas 8.323 komentar YouTube dan 10.124 komentar TikTok. Data tersebut merepresentasikan opini, keluhan, dan tanggapan masyarakat terhadap kondisi serta penanganan banjir.

Tabel 3. Distribusi Data Komentar

Platform	Jumlah Komentar	Persentase
YouTube	8.323	45,11%
TikTok	10.124	54,89%
Total	18.447	100%

Dominasi data TikTok menunjukkan bahwa sumber data penelitian tidak hanya berasal dari *platform* berbasis video panjang, tetapi juga dari *platform* video pendek yang memiliki aktivitas komentar tinggi. Kedua sumber tersebut memberikan variasi teks yang dapat digunakan untuk mengidentifikasi pola keluhan masyarakat.

4.2 Hasil Preprocessing

Komentar media sosial memiliki karakteristik berupa teks tidak terstruktur, penggunaan bahasa informal, variasi penulisan, serta keberadaan simbol dan elemen nonteks. Oleh karena itu, dilakukan *preprocessing* untuk mengurangi *noise* dan menghasilkan teks yang lebih konsisten sebelum proses pemodelan. Tahapan yang digunakan meliputi case folding, cleaning, normalization, tokenization, dan stopword removal. Pada tahap *case folding*, seluruh teks diubah menjadi huruf kecil. Selanjutnya, *cleaning* digunakan untuk menghilangkan elemen yang tidak diperlukan, seperti simbol, emoji, URL, *mention*, angka, dan komentar yang tidak informatif. Normalisasi dilakukan untuk menyeragamkan kata tidak baku dan singkatan, sedangkan *tokenization* memecah teks menjadi unit kata. Tahap terakhir adalah *stopword removal* untuk menghilangkan kata umum yang kurang memberikan kontribusi terhadap pembentukan topik. Hasil *preprocessing* kemudian digunakan sebagai masukan untuk proses pemodelan BERTopic. Tahapan tersebut penting karena kualitas representasi teks pada tahap awal akan memengaruhi proses *embedding*, *clustering*, dan pembentukan topik.

4.3 Hasil Pemodelan Topik

Pemodelan dilakukan menggunakan BERTopic yang menggabungkan *Sentence-BERT*, UMAP, HDBSCAN, dan c-TF-IDF. Dari 18.447 komentar yang telah melalui *preprocessing*, model menghasilkan 268 topik. Seluruh topik kemudian diinterpretasikan berdasarkan kata kunci representatif yang diperoleh dari c-TF-IDF. Untuk memperoleh topik yang paling relevan dengan fokus penelitian, dilakukan *semantic topic matching* menggunakan *Sentence-BERT*, *cosine similarity*, dan *keyword overlap*. Berdasarkan proses tersebut, dipilih 10 topik utama, yaitu, Sebaran Banjir, Kebutuhan Air Bersih, Permintaan Bantuan, Relawan, Masyarakat Terdampak, Akses Bantuan, Penyaluran Bantuan, Kecepatan Penanganan, Status Bencana dan Kritik Pemerintah

Tabel 4. Topik Utama Hasil BERTopic

No.	Topik	Jumlah Komentar
1	Sebaran Banjir	—
2	Kebutuhan Air Bersih	56
3	Permintaan Bantuan	—
4	Relawan	73
5	Masyarakat Terdampak	—
6	Akses Bantuan	14
7	Penyaluran Bantuan	—
8	Kecepatan Penanganan	—
9	Status Bencana	91
10	Kritik Pemerintah	—

Data jumlah komentar yang tersedia secara eksplisit pada hasil skripsi untuk topik-topik tersebut. Status Bencana merupakan topik dengan jumlah komentar terbanyak, yaitu 91 komentar, diikuti Relawan sebanyak 73 komentar dan

Kebutuhan Air Bersih sebanyak 56 komentar. Akses Bantuan memiliki jumlah paling sedikit, yaitu 14 komentar.

4.4 Topik Dominan Keluhan Masyarakat

Hasil pemodelan menunjukkan bahwa keluhan masyarakat tidak hanya berkaitan dengan keberadaan banjir, tetapi mencakup berbagai aspek penanganan bencana. Sepuluh topik yang teridentifikasi dapat dikelompokkan ke dalam beberapa isu utama, yaitu kondisi dan dampak banjir, kebutuhan masyarakat, bantuan dan relawan, kecepatan penanganan, serta evaluasi terhadap pemerintah. Status Bencana menjadi topik dengan jumlah komentar terbesar, yaitu 91 komentar. Topik ini menunjukkan adanya perhatian masyarakat terhadap kondisi dan status bencana yang sedang berlangsung. Selain itu, topik Relawan sebanyak 73 komentar menunjukkan adanya pembahasan mengenai keterlibatan masyarakat atau pihak relawan dalam membantu korban banjir. Sementara itu, Kebutuhan Air Bersih sebanyak 56 komentar menunjukkan bahwa kebutuhan dasar menjadi salah satu perhatian masyarakat dalam kondisi terdampak banjir.

Topik Permintaan Bantuan dan Penyaluran Bantuan menunjukkan bahwa bantuan merupakan salah satu isu penting dalam percakapan masyarakat. Hal tersebut juga diperkuat oleh hasil *word cloud*, di mana kata "bantuan" dan "pertolongan" menjadi kata yang paling dominan. Kondisi ini menunjukkan bahwa kebutuhan dan distribusi bantuan merupakan tema sentral dalam komentar masyarakat. Topik Akses Bantuan memiliki jumlah komentar paling sedikit, yaitu 14 komentar. Meskipun jumlahnya relatif kecil dibandingkan topik lainnya, keberadaannya tetap menunjukkan adanya pembahasan mengenai kesulitan masyarakat dalam memperoleh atau menjangkau bantuan. Topik Kecepatan Penanganan juga menunjukkan perhatian masyarakat terhadap respons terhadap kondisi banjir, terutama ketika penanganan dianggap lambat. Sementara itu, Kritik Pemerintah menunjukkan adanya evaluasi dan ketidakpuasan masyarakat terhadap peran pemerintah dalam penanganan bencana. Hasil *word cloud* memperlihatkan kata "pemerintah" dan "kemana" cukup menonjol, yang mengindikasikan munculnya pertanyaan dan kritik mengenai keberadaan serta peran pemerintah dalam situasi bencana.

Dengan demikian, hasil pemodelan memperlihatkan bahwa keluhan masyarakat memiliki dimensi yang cukup luas. Masyarakat tidak hanya membicarakan kejadian banjir, tetapi juga kebutuhan dasar, bantuan, relawan, akses dan kecepatan penanganan, status bencana, serta kinerja pemerintah.

4.5 Hasil Evaluasi Topic Coherence

Evaluasi kualitas topik dilakukan menggunakan *Normalized Pointwise Mutual Information (NPMI).* Evaluasi dilakukan terhadap 265 topik, dengan mengecualikan *outlier* atau Topic -1. Nilai NPMI keseluruhan diperoleh dengan menghitung rata-rata nilai NPMI dari topik-topik yang dievaluasi.

Tabel 5. Hasil Evaluasi Model

Parameter	Hasil
Jumlah komentar	18.447
Jumlah topik hasil model	268
Topik yang dievaluasi	265
Metrik evaluasi	NPMI
Nilai NPMI	0,1962

Nilai NPMI sebesar 0,1962 menunjukkan tingkat koherensi topik yang cukup atau moderat. Nilai NPMI berada pada rentang -1 hingga 1; nilai yang mendekati 1 menunjukkan koherensi yang tinggi, sedangkan nilai yang mendekati 0 menunjukkan koherensi moderat. Distribusi nilai NPMI setiap topik berada pada rentang sekitar -0,4 hingga 1,0, dengan sebagian besar nilai terkonsentrasi pada rentang 0,0–0,6. Beberapa topik memiliki nilai mendekati 1,0, sedangkan sebagian kecil memiliki nilai negatif. Nilai NPMI yang moderat dapat dipahami berdasarkan karakteristik data penelitian berupa komentar media sosial yang relatif pendek, tidak terstruktur, dan menggunakan bahasa informal. Karakteristik tersebut dapat menyebabkan kata-kata representatif dalam suatu topik tidak selalu muncul secara bersamaan dalam satu komentar.

4.6 Visualisasi Hasil

Visualisasi digunakan untuk membantu interpretasi hasil pemodelan BERTopic. Penelitian menggunakan beberapa bentuk visualisasi, yaitu *Topic Word Score*, *Intertopic Distance Map*, distribusi topik, dan Word Cloud.

4.6.1 *Topic Word Score*

Visualisasi *Topic Word Score* menampilkan kata atau frasa dengan nilai c-TF-IDF tertinggi pada setiap topik. Nilai tersebut menunjukkan tingkat kepentingan suatu kata dalam merepresentasikan topik tertentu. Dari sepuluh topik yang dipilih, Permintaan Bantuan memiliki *Topic Word Score* tertinggi, sekitar 1,50, sedangkan Masyarakat Terdampak memiliki nilai tertinggi sekitar 0,33. Visualisasi ini menunjukkan bahwa kata-kata pada topik Permintaan Bantuan memiliki kemampuan representasi yang lebih kuat dibandingkan topik lainnya. Dengan demikian, isu permintaan bantuan menjadi salah satu tema yang secara semantik cukup kuat dalam dataset.

4.6.2 *Intertopic Distance Map*

Intertopic Distance Map digunakan untuk melihat hubungan antartopik dalam ruang dua dimensi. Setiap lingkaran merepresentasikan suatu topik, sedangkan ukuran lingkaran menunjukkan jumlah dokumen yang termasuk di dalamnya. Jarak antartopik menggambarkan tingkat

kemiripan semantik; topik yang berdekatan memiliki karakteristik pembahasan yang relatif lebih mirip. Visualisasi ini membantu melihat hubungan antara isu yang terbentuk, misalnya kedekatan antara topik bantuan, akses bantuan, dan penyaluran bantuan yang secara konseptual memiliki hubungan dengan proses pemenuhan kebutuhan masyarakat terdampak.

4.6.3 Word Cloud

Hasil *word cloud* memperlihatkan bahwa kata "bantuan" dan "pertolongan" merupakan kata yang paling dominan. Selain itu, kata "pemerintah", "akses", "relawan", "lambat", "air bersih", dan "bencana nasional" juga terlihat menonjol. Temuan tersebut memperkuat hasil pemodelan bahwa percakapan masyarakat banyak berpusat pada bantuan dan pertolongan, kemudian diikuti isu mengenai peran pemerintah, akses bantuan, relawan, kecepatan penanganan, kebutuhan air bersih, dan status bencana.

4.7 Pembahasan

Hasil penelitian menunjukkan bahwa BERTopic mampu mengorganisasi 18.447 komentar masyarakat menjadi kelompok topik yang dapat diinterpretasikan. Dari 268 topik yang terbentuk, sepuluh topik dipilih berdasarkan relevansinya terhadap fokus penelitian. Hasil tersebut menunjukkan bahwa keluhan masyarakat terhadap penanganan banjir memiliki struktur isu yang lebih kompleks daripada sekadar ekspresi positif atau negatif. Temuan yang paling menonjol adalah kuatnya isu bantuan dan pertolongan. Dominasi kata tersebut pada *word cloud*, tingginya *Topic Word Score* pada topik Permintaan Bantuan, serta munculnya topik Kebutuhan Air Bersih, Akses Bantuan, Penyaluran Bantuan, dan Relawan menunjukkan bahwa kebutuhan dan mekanisme bantuan merupakan bagian penting dari percakapan masyarakat.

Di sisi lain, munculnya topik Kecepatan Penanganan dan Kritik Pemerintah menunjukkan bahwa masyarakat juga memberikan evaluasi terhadap respons dan peran pihak berwenang. Hal ini memperlihatkan bahwa komentar media sosial tidak hanya menyampaikan kondisi bencana, tetapi juga menjadi media bagi masyarakat untuk menyampaikan penilaian terhadap proses penanganan. Temuan tersebut memperkuat posisi penelitian dibandingkan pendekatan yang hanya melakukan analisis sentimen. Penelitian ini tidak berhenti pada identifikasi kecenderungan sentimen, tetapi memetakan isu spesifik yang menjadi bahan pembicaraan dan keluhan masyarakat. Dengan demikian, BERTopic memberikan representasi yang lebih terstruktur mengenai dimensi keluhan masyarakat, mulai dari kondisi bencana hingga kebutuhan, bantuan, kecepatan respons, dan evaluasi terhadap pemerintah.

Nilai NPMI sebesar 0,1962 menunjukkan bahwa kualitas koherensi model berada pada tingkat moderat. Meskipun belum menunjukkan koherensi yang tinggi secara keseluruhan, hasil tersebut tetap menunjukkan bahwa model mampu menghasilkan sejumlah topik yang dapat diinterpretasikan. Kondisi ini juga konsisten dengan karakteristik komentar media sosial yang pendek, informal, dan tidak terstruktur sehingga hubungan kemunculan bersama antarkata dalam satu komentar menjadi lebih terbatas.

Secara keseluruhan, hasil penelitian menunjukkan bahwa pemetaan topik dapat mengubah kumpulan komentar masyarakat yang tidak terstruktur menjadi peta isu yang lebih sistematis. Sepuluh topik yang diperoleh memberikan gambaran mengenai aspek-aspek yang paling sering menjadi perhatian masyarakat dalam konteks penanganan banjir, sehingga hasilnya dapat digunakan sebagai informasi pendukung untuk memahami pola keluhan publik dalam komunikasi dan penanganan bencana.

V. KESIMPULAN

Berdasarkan hasil analisis terhadap 18.447 komentar YouTube dan TikTok, metode BERTopic mampu memetakan keluhan masyarakat terhadap penanganan banjir secara sistematis melalui tahapan *preprocessing*, *Sentence-BERT*, UMAP, HDBSCAN, dan c-TF-IDF. Pemodelan menghasilkan 268 topik, kemudian melalui *semantic topic matching* diperoleh 10 topik relevan, yaitu Sebaran Banjir, Kebutuhan Air Bersih, Permintaan Bantuan, Relawan, Masyarakat Terdampak, Akses Bantuan, Penyaluran Bantuan, Kecepatan Penanganan, Status Bencana, dan Kritik Pemerintah. Topik Status Bencana menjadi yang terbanyak dengan 91 komentar, diikuti Relawan (73) dan Kebutuhan Air Bersih (56). Isu bantuan dan pertolongan juga menjadi tema menonjol berdasarkan hasil visualisasi. Evaluasi menggunakan NPMI terhadap 265 topik menghasilkan nilai 0,1962, yang menunjukkan tingkat koherensi moderat. Dengan demikian, BERTopic dapat digunakan untuk mengubah komentar masyarakat yang tidak terstruktur menjadi peta isu yang menggambarkan perhatian publik terhadap kondisi bencana, kebutuhan bantuan, relawan, kecepatan penanganan, dan kinerja pemerintah.

REFERENSI

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, pp. 3713–3744, 2023.

[2] N. Reimers and I. Gurevych, "*Sentence-BERT*: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992.

[3] L. McInnes, J. Healy, and J. Melville, "UMAP: *Uniform Manifold Approximation and Projection* for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[4] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2013, pp. 160–172.
[5] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.
[6] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM)*, 2015, pp. 399–408.
[7] W. Wahyuni *et al*., "Analisis persepsi masyarakat terhadap Program Makan Bergizi Gratis menggunakan BERTopic pada komentar YouTube," 2025.
[8] Aryani *et al*., "Identifikasi topik dominan pada pemberitaan Pemilu 2024 menggunakan BERTopic," 2024.
[9] Erwanda *et al*., "Analisis respons masyarakat terhadap bencana banjir di Aceh menggunakan pendekatan lexicon-based sentiment analysis," 2025.
[10] L. M. Lestari, A. Diniati, P. Alamsyah, and A. N. Salma, "Taxonomy of community response during flood disasters in Jakarta: A communication perspective," *Jurnal Sosioteknologi*, vol. 23, no. 3, pp. 438–449, 2024.
[11] Y. Y. Anggara and N. Arif, "Analisis respons publik dan permodelan topik menggunakan Latent Dirichlet Allocation method (LDA) pada bencana banjir bandang di Sumatera Barat 2024 melalui Twitter," *Indonesian Journal of Environment and Disaster*.
[12] T. Acikara, B. Xia, T. Yigitcanlar, and C. Hon, "Contribution of social media analytics to disaster response effectiveness: A systematic review of the literature," *Sustainability*, vol. 15, no. 11, p. 8860, 2023.
[13] G. A. Yudarwati, I. A. Putranto, and K. M. Delmo, "Examining the Indonesian government's social media use for disaster risk communication," *Asian Journal of Communication*, vol. 32, no. 1, 2022.
[14] K. Guo, M. Guan, and H. Yan, "Utilising social media data to evaluate urban flood impact in data scarce cities," *International Journal of Disaster Risk Reduction*, vol. 93, p. 103780, 2023.
[15] V. V. Mihunov *et al*., "Disaster impacts surveillance from social media with topic modeling and feature extraction: Case of Hurricane Harvey," *International Journal of Disaster Risk Science*, vol. 13, pp. 729–742, 2022.
[16] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022, doi: 10.48550/arXiv.2203.05794.