Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 5 No. 1 (2026) pp: 797-807
P-ISSN: 2986-6790, e-ISSN: 2986-6782

Analisis Topik Percakapan Masyarakat tentang Kebakaran Hutan dan Lahan (Karhutla) di Indonesia Menggunakan BERTopic

Erlina Ali Marpaung, Zulfan Afandi, Iwan Fitrianto Rahmad, Wendi Amsuri Nst, Dea Balqis Intan Pratiwi
Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Potensi Utama, Medan, Indonesia
erlinaalimarpaung@gmai.com, zulfan.computer@gmail.com, fitrianto.iwan@gmail.com, mrs.wendhi@gmail.com, deabakqis@gmail.com

Abstrak

Kebakaran hutan dan lahan (Karhutla) merupakan permasalahan lingkungan yang menimbulkan beragam perhatian dan respons masyarakat di media sosial. Penelitian ini bertujuan mengidentifikasi, mengelompokkan, dan menganalisis topik percakapan masyarakat mengenai Karhutla di Indonesia, khususnya percakapan yang membahas wilayah Kalimantan, melalui komentar dan balasan komentar Instagram menggunakan metode BERTopic. Data dikumpulkan melalui proses scraping dengan bahasa pemrograman Python. Sebanyak 3.207 komentar berhasil diperoleh, kemudian dilakukan pembersihan data sehingga 2.061 komentar digunakan dalam pemodelan. Hasil pemodelan BERTopic menghasilkan 14 topik valid. Dari keseluruhan data yang dianalisis, sebanyak 1.808 komentar berhasil dikelompokkan ke dalam topik, sedangkan 253 komentar dikategorikan sebagai outlier. Hasil analisis menunjukkan bahwa percakapan masyarakat mencakup beberapa isu utama, yaitu kebakaran hutan, dampak kebakaran terhadap hewan, dugaan pembakaran yang disengaja, harapan terhadap turunnya hujan, kritik terhadap pemerintah, serta bentuk kepedulian masyarakat. Untuk memahami karakteristik dan hubungan antartopik, penelitian ini memanfaatkan distribusi topik, kata kunci berbobot class-based TF-IDF (c-TF-IDF), matriks kemiripan, dan dendrogram. Hasil visualisasi memperlihatkan adanya keterkaitan makna di antara beberapa topik, sekaligus menunjukkan variasi fokus pembahasan masyarakat. Temuan ini menggambarkan bahwa percakapan mengenai Karhutla tidak hanya berisi informasi tentang kebakaran, tetapi juga memuat respons, penilaian, dan kepedulian publik terhadap dampaknya. Dengan demikian, BERTopic dapat dimanfaatkan untuk mengidentifikasi dan mengelompokkan topik percakapan media sosial berdasarkan kemiripan semantik teks serta memberikan gambaran mengenai isu-isu Karhutla yang menjadi perhatian masyarakat.

Kata kunci: Karhutla; Instagram; BERTopic; Pemodelan Topik; Analisis Teks.

Abstract

Forest and land fires (Karhutla) are environmental problems that attract public attention and generate diverse responses on social media. This study aims to identify, group, and analyze public discussion topics about Karhutla in Indonesia, particularly discussions referring to the Kalimantan region, through Instagram comments and replies using the BERTopic method. Data were collected through scraping with the Python programming language. A total of 3,207 comments were obtained and processed through data cleaning, resulting in 2,061 comments being used for topic modeling. The BERTopic model identified 14 valid topics. Of the analyzed comments, 1,808 were assigned to topic clusters, while 253 were categorized as outliers. The analysis revealed several major themes in public discussions, including forest fires, the impacts of fires on animals, alleged deliberate burning, hopes for rainfall, criticism of the government, and expressions of public concern. To examine the characteristics and relationships among the topics, the study used topic distribution, class-based TF-IDF (c-TF-IDF) keywords, a similarity matrix, and a hierarchical clustering dendrogram. The visualizations indicate semantic relationships among several topics while also illustrating variations in the focus of public discussions. These findings show that discussions about Karhutla extend beyond the fires themselves to include public responses, evaluations, and concern about their impacts. Therefore, BERTopic can be used to identify and group social media discussion topics based on semantic similarities in text and to provide an overview of Karhutla-related issues that attract public attention.

Keywords: Forest and Land Fires; Instagram; BERTopic; Topic Modeling; Text Analysis.

1. Pendahuluan

Kebakaran hutan dan lahan (Karhutla) merupakan salah satu permasalahan lingkungan yang masih menjadi perhatian di Indonesia. Peristiwa ini tidak hanya menyebabkan kerusakan ekosistem, tetapi juga menimbulkan pencemaran udara akibat asap yang berdampak terhadap kesehatan, aktivitas masyarakat, serta kondisi sosial dan

ekonomi. Permasalahan Karhutla juga memunculkan berbagai tanggapan dari masyarakat, mulai dari pembahasan mengenai penyebab kebakaran, dampak terhadap lingkungan, kondisi masyarakat terdampak, hingga kritik terhadap kebijakan dan penanganan pemerintah. Seiring dengan perkembangan teknologi informasi, media sosial menjadi salah satu sarana bagi masyarakat untuk menyampaikan pendapat, membagikan informasi, dan memberikan tanggapan terhadap berbagai peristiwa. Salah satu platform yang banyak dimanfaatkan untuk keperluan tersebut adalah Instagram. Melalui komentar dan balasan komentar, masyarakat dapat menyampaikan berbagai pandangan serta informasi mengenai permasalahan Karhutla. Dengan demikian, aktivitas komunikasi di Instagram dapat dimanfaatkan sebagai sumber data untuk memahami isu-isu yang berkembang dalam percakapan masyarakat mengenai Karhutla.

Permasalahan kebakaran hutan dan lahan di Indonesia bukan merupakan peristiwa yang baru terjadi, melainkan permasalahan lingkungan yang telah berulang dalam beberapa tahun terakhir. Arisman (2020) melalui penelitian mengenai tren kebakaran hutan dan lahan di Indonesia periode 2015–2019 menunjukkan bahwa kejadian Karhutla terus berlangsung selama periode pengamatan dan memiliki pola peningkatan pada musim kemarau. Selanjutnya, Nurhayati et al. (2021) mengkaji dinamika kebakaran hutan dan lahan gambut di Provinsi Sumatera Selatan selama periode 2015–2019. Penelitian tersebut menunjukkan bahwa Kabupaten Ogan Komering Ilir merupakan salah satu wilayah yang mengalami kebakaran berulang, dengan tahun 2015 sebagai periode dengan jumlah titik panas tertinggi dalam rentang pengamatan. Temuan tersebut memperlihatkan bahwa Karhutla merupakan permasalahan yang berulang dan berkaitan dengan kondisi lingkungan serta karakteristik lahan, khususnya lahan gambut.

Kejadian Karhutla juga ditemukan di wilayah Sumatera Selatan pada periode berikutnya. Zaskia et al. (2023) melalui penelitian mengenai pengaruh curah hujan terhadap kejadian kebakaran hutan dan lahan di Kabupaten Banyuasin, Sumatera Selatan, membahas hubungan kondisi curah hujan dengan kejadian kebakaran selama periode pengamatan. Selain itu, Hamdi et al. (2025) mengkaji penyebaran kabut asap dari kebakaran lahan gambut di Sumatera Selatan pada 2023. Penelitian tersebut menunjukkan bahwa asap kebakaran dapat menyebar hingga wilayah perkotaan dan mengganggu jarak pandang. Berbagai penelitian tersebut memperlihatkan bahwa Karhutla telah terjadi secara berulang di Indonesia, termasuk di wilayah Sumatera, serta menimbulkan dampak terhadap lingkungan dan kehidupan masyarakat. Kondisi ini menunjukkan pentingnya mengkaji permasalahan Karhutla tidak hanya melalui data kejadian kebakaran, tetapi juga melalui respons dan pembahasan masyarakat terhadap peristiwa tersebut.

Pemanfaatan Instagram sebagai sumber informasi dalam penelitian mengenai isu lingkungan telah dilakukan oleh Ardhika et al. (2025) melalui penelitian berjudul "Social Media for Wildlife Conservation Campaign in South Sumatra, Indonesia". Penelitian tersebut menganalisis 470 unggahan Instagram dari Balai Konservasi Sumber Daya Alam (BKSDA) Sumatera Selatan dan Belantara Foundation selama tahun 2022. Hasil penelitian menunjukkan bahwa Instagram dimanfaatkan untuk menyampaikan informasi mengenai konservasi satwa liar, kegiatan pelestarian lingkungan, serta edukasi kepada masyarakat. Penelitian tersebut memperlihatkan bahwa Instagram dapat digunakan sebagai media komunikasi untuk menyampaikan informasi dan meningkatkan perhatian masyarakat terhadap isu lingkungan.

Selanjutnya, Yao dan Pearce (2025) melalui penelitian berjudul "The Visual Vernacular of Climate Change on Instagram: How Modal Convergence Between Image and Text Is Changing the Representation of Climate Solutions" menganalisis representasi isu perubahan iklim melalui konten Instagram. Penelitian tersebut menggunakan pendekatan kuantitatif dan kualitatif untuk mengidentifikasi tema serta bentuk komunikasi visual yang berkaitan dengan perubahan iklim. Hasil penelitian menunjukkan bahwa konten Instagram dapat menggambarkan berbagai perspektif mengenai perubahan iklim melalui kombinasi gambar dan teks. Selain itu, Khairani et al. (2026) dalam penelitian berjudul "Peran Media Sosial Instagram @BPBDPROVINSISUMUT dan @INFOBMKGSUMUT dalam Penyebaran Informasi Bencana Banjir di Sumatera Utara" membahas pemanfaatan Instagram dalam penyebaran informasi bencana. Penelitian tersebut menunjukkan bahwa Instagram berperan sebagai media penyampaian informasi sebelum, selama, dan setelah bencana melalui unggahan informasi, infografis, serta dokumentasi lapangan. Berdasarkan penelitian-penelitian tersebut, Instagram memiliki potensi untuk dimanfaatkan sebagai sumber data dalam mengkaji komunikasi publik, penyebaran informasi, dan perhatian masyarakat terhadap permasalahan lingkungan maupun bencana.

Selain pemanfaatan Instagram sebagai media komunikasi publik, penelitian mengenai analisis percakapan masyarakat di media sosial juga berkembang melalui penerapan metode topic modeling. Salah satu metode yang dapat digunakan adalah BERTopic. Anugrah dan Agatha (2025) melalui penelitian berjudul "Utilizing

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts" menganalisis sebanyak 3.075 komentar pada akun Instagram Badan Pusat Statistik (BPS). Penelitian tersebut menggunakan IndoBERT untuk klasifikasi sentimen dan BERTopic untuk mengidentifikasi kelompok topik dalam komentar masyarakat. Hasil penelitian menunjukkan bahwa BERTopic dapat digunakan untuk menemukan berbagai kelompok pembahasan berdasarkan komentar yang disampaikan pengguna Instagram. Selanjutnya, Lie et al. (2026) melalui penelitian berjudul "Topic Modeling and Text Classification of Public Comments Towards BPJS Kesehatan Using BERTopic and IndoBERT" menggunakan data komentar Instagram BPJS Kesehatan periode 2021–2022. Penelitian tersebut menggabungkan IndoBERT dan BERTopic untuk mengklasifikasikan serta mengidentifikasi topik utama dalam komentar masyarakat. Hasil pemodelan menghasilkan beberapa kelompok pembahasan yang berkaitan dengan keluhan dan pertanyaan masyarakat mengenai pelayanan BPJS Kesehatan. Kedua penelitian tersebut menunjukkan bahwa BERTopic dapat diterapkan untuk menganalisis komentar Instagram dan mengelompokkan pembahasan masyarakat berdasarkan kemiripan konteks serta makna teks.

Secara metodologis, Grootendorst (2022) memperkenalkan BERTopic melalui penelitian berjudul "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure". Metode ini menggabungkan representasi teks berbasis transformer, pengelompokan dokumen, dan pendekatan class-based TF-IDF untuk menghasilkan kelompok topik berdasarkan kemiripan makna antardokumen. Berbeda dengan pendekatan topic modeling yang hanya mengandalkan frekuensi kemunculan kata, BERTopic memanfaatkan representasi semantik untuk mengidentifikasi hubungan makna antarteks. Kemampuan tersebut memungkinkan metode ini digunakan untuk menganalisis kumpulan dokumen yang memiliki variasi kata dan konteks pembahasan.

Penerapan BERTopic dalam kajian isu lingkungan ditunjukkan oleh Gökçimen dan Daş (2024) melalui penelitian berjudul "Exploring Climate Change Discourse on Social Media and Blogs Using a Topic Modeling Analysis". Penelitian tersebut menggunakan BERTopic dan LDA untuk mengidentifikasi serta memvisualisasikan topik pembahasan mengenai perubahan iklim dari data media sosial dan blog. Hasil penelitian menunjukkan bahwa BERTopic dapat digunakan untuk mengidentifikasi kelompok pembahasan berdasarkan kemiripan semantik serta memberikan gambaran mengenai persepsi masyarakat terhadap isu lingkungan. Sementara itu, penerapan BERTopic pada data media sosial yang berkaitan dengan bencana juga dilakukan oleh Murthy et al. (2025) melalui penelitian berjudul "Examining Hurricane-Related Social Media Topics Longitudinally and at Scale: A Transformer-Based Approach". Penelitian tersebut menganalisis sekitar 1,7 juta unggahan media sosial mengenai empat peristiwa badai di Amerika Serikat selama tujuh tahun. Dengan menggunakan BERTopic, penelitian tersebut berhasil mengidentifikasi berbagai kelompok pembahasan, seperti bantuan bencana, kekhawatiran masyarakat, kondisi cuaca, dan proses pemulihan. Penelitian tersebut memperlihatkan bahwa BERTopic dapat diterapkan untuk mengeksplorasi percakapan masyarakat dalam jumlah besar serta mengelompokkan informasi berdasarkan kesamaan konteks pembahasan.

Berdasarkan penelitian terdahulu, pemanfaatan Instagram telah diterapkan untuk mengkaji komunikasi publik, penyebaran informasi lingkungan, dan respons masyarakat terhadap berbagai peristiwa bencana. Di sisi lain, penerapan BERTopic telah menunjukkan kemampuannya dalam mengidentifikasi dan mengelompokkan topik percakapan masyarakat berdasarkan kemiripan makna teks. Meskipun demikian, masih terdapat peluang penelitian untuk mengkaji secara khusus pola percakapan masyarakat mengenai kebakaran hutan dan lahan (Karhutla) di Indonesia melalui data Instagram menggunakan metode BERTopic. Penelitian mengenai Karhutla di media sosial sebelumnya telah membahas keterlibatan masyarakat, jaringan komunikasi, dan respons publik terhadap bencana. Namun, eksplorasi topik percakapan masyarakat yang berfokus pada pengelompokan berdasarkan kemiripan semantik dari data caption, komentar, dan balasan komentar Instagram masih perlu dikembangkan. Oleh karena itu, diperlukan pendekatan yang tidak hanya mengidentifikasi frekuensi kemunculan kata, tetapi juga mampu mengungkap kelompok pembahasan berdasarkan hubungan makna antarteks.

Kebaruan (novelty) penelitian ini terletak pada penerapan metode BERTopic untuk mengidentifikasi dan memetakan topik percakapan masyarakat mengenai Karhutla di Indonesia berdasarkan data Instagram. Data yang digunakan mencakup komentar dan balasan komentar sehingga analisis tidak hanya berfokus pada informasi yang disampaikan melalui unggahan, tetapi juga mencakup tanggapan dan interaksi masyarakat. Berbeda dengan penelitian terdahulu yang lebih berfokus pada komunikasi lingkungan, penyebaran informasi bencana, klasifikasi sentimen, maupun keterlibatan masyarakat, penelitian ini menitikberatkan pada pengelompokan percakapan berdasarkan representasi semantik. Melalui pendekatan tersebut, penelitian diharapkan dapat mengungkap berbagai kelompok isu yang menjadi perhatian masyarakat, seperti penyebab kebakaran, dampak asap, kerusakan lingkungan, kondisi masyarakat terdampak, serta respons dan penanganan pemerintah berdasarkan topik yang terbentuk dari hasil pemodelan.

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengidentifikasi, mengelompokkan, dan menganalisis topik-topik utama dalam percakapan masyarakat mengenai kebakaran hutan dan lahan (Karhutla) di Indonesia, khususnya percakapan yang membahas wilayah Kalimantan, menggunakan metode BERTopic dengan memanfaatkan data Instagram berupa komentar dan balasan komentar. Selain itu, penelitian ini bertujuan untuk mengetahui kecenderungan pembahasan masyarakat berdasarkan kelompok topik yang dihasilkan, sehingga dapat memberikan gambaran mengenai isu-isu yang menjadi perhatian publik, termasuk pembahasan mengenai kebakaran, pihak yang dianggap bertanggung jawab, doa dan harapan, kritik, serta respons empati dan kesedihan terhadap permasalahan Karhutla di wilayah Kalimantan.

2. Metode Penelitian

2.1 Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif. Pendekatan ini digunakan untuk mengidentifikasi dan menggambarkan berbagai topik yang muncul dalam percakapan masyarakat mengenai kebakaran hutan dan lahan (Karhutla) di Indonesia. Analisis dilakukan dengan memanfaatkan data teks dari Instagram yang diolah menggunakan metode BERTopic untuk menemukan kelompok pembahasan berdasarkan kemiripan makna antardokumen. Menurut Grootendorst (2022), BERTopic merupakan metode pemodelan topik yang memanfaatkan representasi semantik berbasis transformer untuk mengelompokkan dokumen berdasarkan kesamaan konteks.

2.2 Sumber dan Pengumpulan Data

Data penelitian merupakan data sekunder yang diperoleh dari platform Instagram melalui teknik *data scraping* menggunakan bahasa pemrograman Python. Salah satu sumber data yang digunakan adalah unggahan Instagram pada tautan *https://www.instagram.com/reel/DcRH8H9T914/*. Pengumpulan data dilakukan dengan menggunakan kata kunci yang berkaitan dengan Karhutla, seperti kebakaran hutan, kebakaran lahan, Karhutla, dan kabut asap. Data yang dikumpulkan meliputi teks komentar utama dan balasan komentar (*replies*) dari pengguna pada unggahan yang membahas Karhutla, khususnya wilayah Kalimantan. Komentar dan balasan yang relevan digunakan sebagai bahan analisis, sedangkan data kosong, duplikat, dan tidak berkaitan dengan penelitian tidak disertakan. Seluruh data selanjutnya diolah menggunakan Python untuk tahap pembersihan data dan pemodelan BERTopic. Identitas pengguna disamarkan untuk menjaga privasi dalam penyajian hasil penelitian.

2.3 Pembersihan Data (Data Cleaning)

Tahap pembersihan data dilakukan terhadap 3.207 komentar Instagram untuk memastikan data sesuai dengan kebutuhan penelitian. Proses ini meliputi penghapusan komentar duplikat, teks kosong, URL, simbol, karakter khusus, serta komentar yang tidak berkaitan dengan Karhutla. Tahapan ini bertujuan mengurangi data yang tidak relevan agar proses analisis topik menggunakan BERTopic dapat berjalan lebih optimal (Uthirapathy & Sandanam, 2023). Setelah proses pembersihan, diperoleh 2.061 komentar yang digunakan dalam penelitian, sedangkan 1.146 komentar lainnya tidak digunakan karena tidak memenuhi kriteria pengolahan data.

2.4 Prapemrosesan Teks (Text Preprocessing)

Setelah proses pembersihan, data dilanjutkan ke tahap prapemrosesan teks untuk menyeragamkan dan menyederhanakan bentuk tulisan. Tahapan yang dilakukan meliputi:

1. Case folding, mengubah seluruh huruf menjadi huruf kecil.
2. Cleansing, menghapus karakter atau simbol yang tidak diperlukan.
3. Tokenizing, memisahkan kalimat menjadi kata-kata.
4. Normalization, mengubah kata tidak baku menjadi bentuk baku.
5. Stopword removal, menghapus kata umum yang kurang memberikan informasi.

Prapemrosesan teks bertujuan meningkatkan kualitas data sebelum memasuki tahap pemodelan sehingga informasi yang relevan dapat diproses secara lebih optimal (Uthirapathy & Sandanam, 2023).

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2.5 Pemodelan Topik Menggunakan BERTopic

Tahap pemodelan topik merupakan proses utama dalam penelitian ini. Metode BERTopic digunakan untuk mengelompokkan dokumen berdasarkan kemiripan makna sehingga menghasilkan topik-topik yang menggambarkan pembahasan masyarakat mengenai Karhutla. Proses pemodelan dilakukan melalui beberapa tahapan, yaitu:

1. Document Embedding mengubah teks menjadi representasi numerik menggunakan Sentence Transformer.
2. Dimensionality Reduction mengurangi dimensi data menggunakan UMAP.
3. Clustering mengelompokkan dokumen berdasarkan kemiripan menggunakan HDBSCAN.
4. Topic Representation menentukan kata-kata utama setiap topik menggunakan Class-Based TF-IDF (c-TF-IDF).

Menurut Grootendorst (2022), BERTopic menggabungkan representasi teks berbasis transformer, proses pengelompokan dokumen, dan c-TF-IDF untuk menghasilkan representasi topik yang lebih koheren.

2.6 Analisis dan Interpretasi Topik

Setelah pemodelan selesai, hasil pengelompokan dianalisis berdasarkan kata-kata dominan, jumlah dokumen, dan contoh unggahan pada setiap topik. Selanjutnya, setiap topik diberikan label sesuai konteks pembahasannya untuk mengetahui isu-isu yang menjadi perhatian masyarakat mengenai Karhutla. Pendekatan ini sejalan dengan penelitian Gökçimen dan Daş (2024) yang menggunakan BERTopic untuk mengidentifikasi dan memvisualisasikan topik pembahasan mengenai perubahan iklim.

2.7 Visualisasi Hasil Penelitian

Tahap terakhir adalah menyajikan hasil pemodelan topik dalam bentuk visualisasi agar pola percakapan masyarakat lebih mudah dipahami. Visualisasi yang digunakan meliputi distribusi jumlah dokumen pada setiap topik, kata-kata dominan, serta hubungan atau kemiripan antartopik. Hasil visualisasi kemudian digunakan untuk menjelaskan topik yang paling banyak dibicarakan dan keterkaitan isu-isu yang muncul dalam percakapan masyarakat mengenai Karhutla. Pendekatan visualisasi topik juga digunakan dalam penelitian Gökçimen dan Daş (2024) untuk memperlihatkan kelompok pembahasan mengenai perubahan iklim dari data media sosial dan blog.

3. Hasil dan Pembahasan

3.1 Gambaran Umum Data Penelitian

Penelitian ini menggunakan data komentar Instagram yang berkaitan dengan isu kebakaran hutan dan lahan (Karhutla) di Indonesia, dengan fokus pembahasan pada peristiwa Karhutla di wilayah Kalimantan. Data diperoleh melalui proses pengumpulan komentar pada unggahan Instagram yang membahas peristiwa Karhutla. Berdasarkan hasil pengumpulan data, diperoleh sebanyak 3.207 komentar sebagai data awal penelitian. Selanjutnya, dilakukan proses pembersihan data (*data cleaning*) dengan menghapus komentar kosong dan komentar duplikat agar data yang digunakan tidak mengandung informasi berulang serta memiliki isi teks yang dapat dianalisis. Setelah proses pembersihan, diperoleh sebanyak 2.061 komentar yang digunakan untuk tahap *preprocessing* dan pemodelan BERTopic.

Proses pembersihan tersebut sejalan dengan penelitian Mariwy et al. (2025) yang menerapkan tahapan preprocessing berupa penghapusan data duplikat dan pembersihan teks sebelum melakukan analisis komentar media sosial. Tahapan ini diperlukan agar data yang digunakan lebih terstruktur dan dapat mendukung proses analisis secara optimal.

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

3.2 Tahap Preprocessing dan Pemodelan Topik

A. Tahap Preprocessing Teks

Untuk menghasilkan data yang bersih, terstruktur, dan siap digunakan dalam proses pemodelan, dilakukan tiga tahapan utama preprocessing teks, yaitu cleaning, normalisasi, dan stopword removal. Setiap tahapan memiliki fungsi masing-masing dalam mengolah data mentah menjadi teks yang lebih konsisten dan mudah dianalisis. Pada tahap stopword removal, digunakan daftar sebanyak 121 kata stopword untuk menghapus kata-kata umum yang kurang memberikan informasi penting dalam pembentukan topik. Contoh hasil setiap tahapan ditampilkan pada Gambar 1 sampai dengan Gambar 3.

	komentar	cleaning
0	semoga Allah azab yang bakar hutan dengan seng...	semoga allah azab yang bakar hutan dengan seng...
1	Kalau ada oknum yang sengaja membakar hutan di...	kalau ada oknum yang sengaja membakar hutan di...
2	Doakan kami Kalimantan agar turun hujan..	doakan kami kalimantan agar turun hujan
3	gue di titik nangisin keadaan negara yaAllah	gue di titik nangisin keadaan negara yaallah
4	Semoga Allah segera turunkan hujan di Kalimant...	semoga allah segera turunkan hujan di kalimantan
5	Jika emang sengaja dibakar, semoga membusuk da...	jika emang sengaja dibakar semoga membusuk dan...
6	Hutan tidak kebakaran. Hutan Kalimantan sengaj...	hutan tidak kebakaran hutan kalimantan sengaja...
7	Pray for Indonesia i'm from canadaCA	pray for indonesia i m from canada
8	Sakit bgt rasanya dan hancur	sakit bgt rasanya dan hancur
9	Save Indonesia' Forest	save indonesia forest

Gambar 1. Proses Cleaning Komentar

Gambar 1 menunjukkan proses cleaning atau pembersihan teks, yaitu penghapusan emoji, simbol, serta karakter non-huruf yang tidak diperlukan dalam analisis. Pada baris ke-2 dan ke-7, data asli masih mengandung emoji. Setelah proses cleaning, emoji tersebut dihapus sehingga teks menjadi lebih bersih, misalnya "doakan kami kalimantan agar turun hujan". Tahap ini bertujuan mengurangi karakter yang tidak relevan agar teks lebih konsisten dan siap diolah pada tahap berikutnya.

	cleaning	normalisasi
0	semoga allah azab yang bakar hutan dengan seng...	semoga allah azab yang bakar hutan dengan seng...
1	kalau ada oknum yang sengaja membakar hutan di...	kalau ada oknum yang sengaja membakar hutan di...
2	doakan kami kalimantan agar turun hujan	doakan kami kalimantan agar turun hujan
3	gue di titik nangisin keadaan negara yaallah	gue di titik nangisin keadaan negara yaallah
4	semoga allah segera turunkan hujan di kalimantan	semoga allah segera turunkan hujan di kalimantan
5	jika emang sengaja dibakar semoga membusuk dan...	jika emang sengaja dibakar semoga membusuk dan...
6	hutan tidak kebakaran hutan kalimantan sengaja...	hutan tidak kebakaran hutan kalimantan sengaja...
7	pray for indonesia i m from canada	pray for indonesia i m from canada
8	sakit bgt rasanya dan hancur	sakit banget rasanya dan hancur
9	save indonesia forest	save indonesia forest

Gambar 2. Proses Normalisasi Kata

Tahap normalisasi dilakukan untuk mengubah kata tidak baku, singkatan, dan bahasa informal menjadi bentuk yang lebih baku dan mudah dipahami. Pada Gambar 2, kalimat "sakit bgt rasanya" diubah menjadi "sakit banget rasanya" dengan mengganti singkatan "bgt" menjadi "banget". Tahap ini penting karena komentar Instagram banyak menggunakan bahasa sehari-hari dan singkatan. Normalisasi membantu menyeragamkan variasi penulisan sehingga kualitas teks menjadi lebih baik untuk analisis.

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



	komentar	normalisasi	preprocessing
0	semoga Allah azab yang bakar hutan dengan seng...	semoga allah azab yang bakar hutan dengan seng...	semoga allah azab bakar hutan sengaja semoga a...
1	Kalau ada oknum yang sengaja membakar hutan di...	kalau ada oknum yang sengaja membakar hutan di...	kalau oknum sengaja membakar hutan kalimantan ...
2	Doakan kami Kalimantan agar turun hujan..	doakan kami kalimantan agar turun hujan	doakan kalimantan turun hujan
3	gue di titik nangisin keadaan negara yaAllah	gue di titik nangisin keadaan negara yaallah	gue titik nangisin keadaan negara yaallah
4	Semoga Allah segera turunkan hujan di Kalimant...	semoga allah segera turunkan hujan di kalimantan	semoga allah segera turunkan hujan kalimantan
5	Jika emang sengaja dibakar, semoga membusuk da...	jika emang sengaja dibakar semoga membusuk dan...	emang sengaja dibakar semoga membusuk kekal di...
6	Hutan tidak kebakaran. Hutan Kalimantan sengaj...	hutan tidak kebakaran hutan kalimantan sengaja...	hutan tidak kebakaran hutan kalimantan sengaja...
7	Pray for Indonesia /m from canadaca	pray for indonesia i m from canada	pray for indonesia i m from canada
8	Sakit bgt rasanya dan hancur	sakit banget rasanya dan hancur	sakit banget rasanya hancur
9	Save Indonesia' Forest	save indonesia forest	save indonesia forest

Activate Windows

Gambar 3. Proses Hapus Stopword (Preprocessing Akhir)

Tahap terakhir adalah stopword removal, yaitu penghapusan kata-kata umum yang dinilai kurang berkontribusi terhadap pembentukan topik, seperti “yang”, “dengan”, “di”, dan “dan”. Sebagai contoh, kalimat “semoga Allah azab yang bakar hutan dengan sengaja” menjadi “semoga Allah azab bakar hutan sengaja”. Proses ini membuat teks lebih ringkas dan menonjolkan kata-kata yang dianggap penting. Hasil akhir preprocessing kemudian digunakan sebagai masukan dalam pemodelan BERTopic.

B. Hasil Pemodelan Topik BERTopic



	Topic	Count	Name	Representation	Representative_Docs
0	-1	253	-1_allah_saudara_karhutla_semua	[allah, saudara, karhutla, semua, sakit, semog...	[dzalimmmmm demi allah, hidup cuma sekali tolo...
1	0	324	0_hutan_bakar_hewan_membakar	[hutan, bakar, hewan, membakar, sengaja, kebak...	[siapa saja yang dengan sengaja dan tidak seng...
2	1	204	1_allah_semoga_lekas_membaik	[allah, semoga, lekas, membaik, lbt, semoga l...	[semoga lekas membaik ya robb, laa ilaaha illa...
3	2	141	2_hujan_turun_turunkan_turunkan hujan	[hujan, turun, turunkan, turunkan hujan, allah...	[ya allah turunkan hujan di kalimantan, astaqf...
4	3	109	3_pemerintah_negara_tidak_presiden	[pemerintah, negara, tidak, presiden, to, the,...	[pemerintah kita kemana, pemerintah masih foku...
5	4	100	4_indonesia_for_kalimantan_pulih	[indonesia, for, kalimantan, pulih, pray, pray...	[indonesia borneo free borneo, indonesia was b...
6	5	82	5_allah_engkau_orang_semua	[allah, engkau, orang, semua, mu, tuhan, tidak...	[tidak ada yang lebih berkuasa selain engkau y...
7	6	52	6_banget_hati banget_hati_nangis	[banget, hati banget, hati, nangis, sakit, sak...	[huhu sakit hati banget liatnya, sakit hati ba...
8	7	38	7_sedih_sedih banget_banget_so sad	[sedih, sedih banget, banget, so sad, sad, so,...	[hah astagaaa yaallah sedih banget, sedih bang...
9	8	29	8_tidak_tidak tega_tega_liatnya	[tidak, tidak tega, tega, liatnya, sanggup, ti...	[ya allah sedih tidak tega liatnya, ya allah t...
10	9	27	9_pray_pray for_for_for kalimantan	[pray, pray for, for, for kalimantan, kalimant...	[pray for kalimantan, pray for kalimantan, pra...
11	10	27	10_asap_paru_cepat_dunia	[asap, paru, cepat, dunia, segar, asaprnya, uda...	[paru paru dunia sekejap jadi, semoga situasi ...
12	11	24	11_sakit_sakit banget_liatnya_rasanya	[sakit, sakit banget, liatnya, rasanya, banget...	[sakit banget liatnya, sakit banget liatjya, s...
13	12	23	12_lihat_lihatnya_lihat nya_nya allah	[lihat, lihatnya, lihat nya, nya allah, banget...	[ya allah sakit hati sekali lihat nya, sedih l...
14	13	20	13_manusia_dunia_siapa_dunia udah	[manusia, dunia, siapa, dunia udah, hati, oran...	[siapa sih ini biang keladinya astaghfirullah ...

Activ

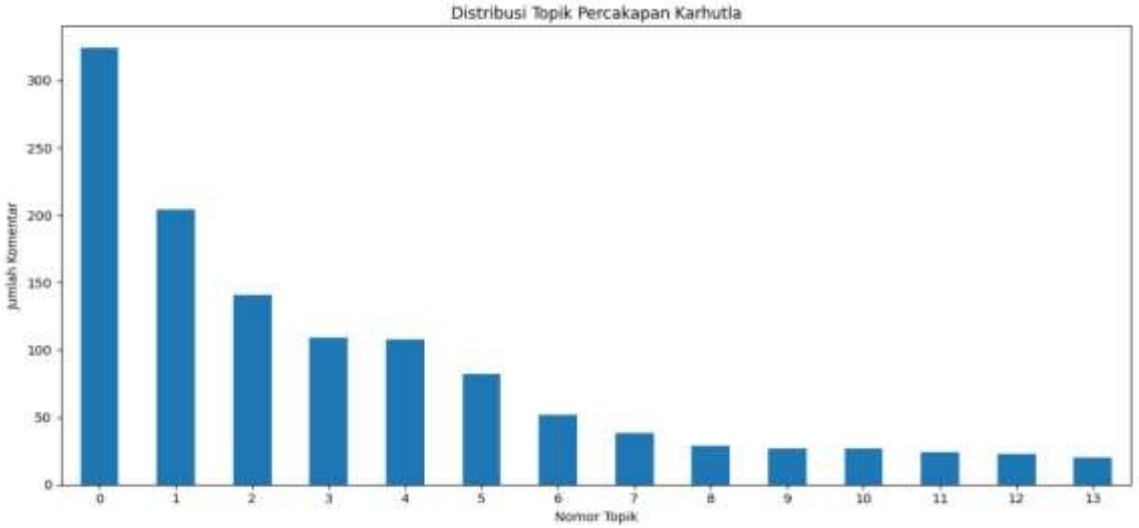
Gambar 4. Informasi Topik yang Terbentuk

Pemodelan BERTopic terhadap 2.061 komentar menghasilkan 14 topik valid, yaitu Topik 0 sampai Topik 13, serta 253 komentar yang dikategorikan sebagai *outlier* (Topik -1). Dengan demikian, sebanyak 1.808 komentar (87,72%) berhasil dikelompokkan ke dalam 14 topik, sedangkan 253 komentar (12,28%) tidak terkelompokkan ke topik tertentu. *Outlier* merupakan komentar yang tidak memiliki kedekatan semantik yang cukup kuat dengan kelompok topik yang terbentuk. Topik 0 menjadi topik terbesar dengan jumlah 324 komentar. Kualitas topik dievaluasi berdasarkan keterkaitan kata kunci dengan tema yang terbentuk, keterwakilan isi komentar, serta kemiripan antartopik melalui visualisasi matriks kemiripan dan hierarki pengelompokan.

Hasil tersebut sejalan dengan mekanisme BERTopic yang menggabungkan representasi dokumen berbasis embedding, proses pengelompokan (clustering), dan representasi topik menggunakan class-based TF-IDF atau c-TF-IDF. Dengan pendekatan ini, pengelompokan mempertimbangkan kedekatan makna antardokumen, bukan hanya kesamaan frekuensi kata (Grootendorst, 2022).

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

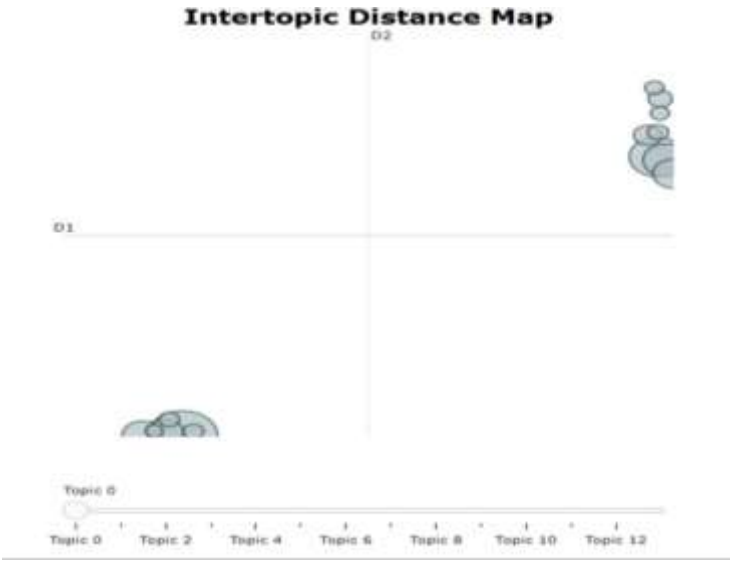
C. Distribusi Topik Percakapan Karhutla



Gambar 5. Grafik Distribusi Jumlah Komentar per Topik

Grafik pada Gambar 5 menunjukkan distribusi topik yang tidak merata (long-tail distribution). Topik 0, yang berkaitan dengan kata "hutan", "bakar", dan "hewan", menjadi topik paling dominan, diikuti Topik 1 yang berkaitan dengan doa dan harapan agar kondisi segera membaik. Jumlah komentar mengalami penurunan cukup tajam setelah Topik 5. Pola ini menunjukkan bahwa percakapan publik lebih banyak terpusat pada sekitar lima hingga enam isu utama, terutama penyebab kebakaran hutan dan harapan masyarakat agar segera turun hujan. Topik lainnya memperoleh perhatian dalam jumlah yang relatif lebih kecil.

D. Visualisasi Jarak Antar Topik (Intertopic Distance Map)



Gambar 6. Intertopic Distance Map

Visualisasi pada Gambar 6 memperlihatkan kedekatan semantik antar-topik dalam ruang dua dimensi (D1 dan D2). Secara umum, terlihat dua kelompok besar. Kelompok di sisi kanan atas mencakup Topik 0–5 yang cenderung membahas isu substantif, seperti kebakaran hutan, pemerintah, dan kondisi Indonesia.

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Sementara itu, kelompok di sisi kiri bawah mencakup Topik 6–13 yang lebih banyak menggambarkan respons emosional masyarakat, seperti rasa sedih, sakit hati, dan rasa tidak tega. Jarak antarkelompok mengindikasikan perbedaan fokus semantik antara komentar yang membahas persoalan Karhutla dan komentar yang mengekspresikan perasaan atau kepedulian. Interpretasi ini perlu dibaca sebagai kecenderungan berdasarkan posisi visualisasi, bukan sebagai pemisahan mutlak antara isu rasional dan emosional.

E. Representasi Kata Kunci Tiap Topik (Topic Word Scores)

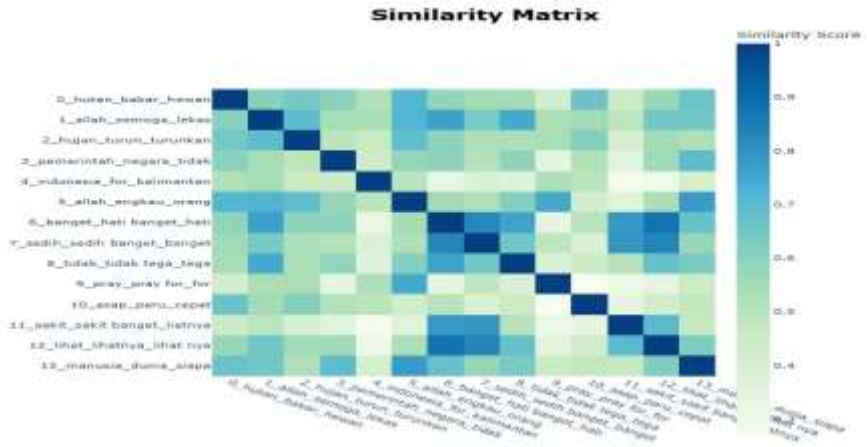


Gambar 7. Skor Kata per Topik (c-TF-IDF)

Setiap topik direpresentasikan oleh kata-kata dengan nilai c-TF-IDF tertinggi yang menunjukkan karakteristik utama topik tersebut. Topik 0, dengan kata kunci “hutan”, “bakar”, “hewan”, “membakar”, dan “sengaja”, menggambarkan kemarahan terhadap tindakan pembakaran hutan. Topik 2, yang memuat kata “hujan”, “turun”, “turungkan”, dan “Allah”, menunjukkan doa serta harapan masyarakat agar hujan turun. Topik 3, dengan kata kunci “pemerintah”, “negara”, dan “presiden”, mencerminkan kritik terhadap pemerintah. Sementara itu, Topik 6 dan 7, yang memuat kata seperti “banget”, “hati”, “nangis”, dan “sedih”, menggambarkan ekspresi kesedihan mendalam.

Nilai c-TF-IDF yang relatif tinggi, sekitar 0,1–0,3 pada kata-kata yang ditampilkan, menunjukkan bahwa kata tersebut memiliki kekhasan dalam merepresentasikan topik terkait dibandingkan dengan distribusi kata pada topik lainnya. Oleh karena itu, interpretasi topik dilakukan dengan memperhatikan kumpulan kata kunci secara bersamaan dan konteks komentar, bukan berdasarkan satu kata secara terpisah.

F. Similarity Matrix Antar Topik

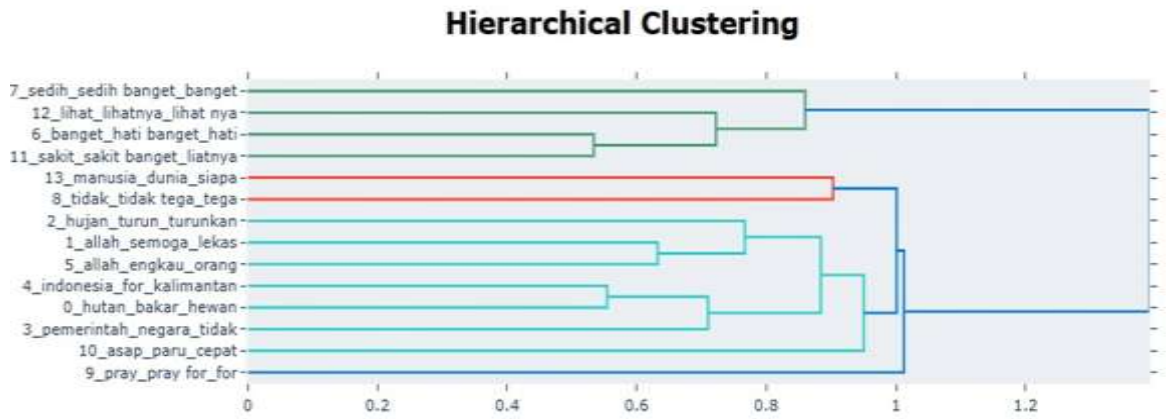


Gambar 8. Similarity Matrix

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Matriks pada Gambar 8 menunjukkan tingkat kemiripan kosinus (cosine similarity) antar-topik dengan rentang nilai yang ditampilkan sekitar 0,4–1,0. Warna biru tua merepresentasikan tingkat kemiripan yang lebih tinggi. Terlihat bahwa Topik 6, 7, 11, dan 12 memiliki kemiripan semantik yang sangat tinggi, sekitar 0,85–0,9. Keempat topik tersebut memuat ungkapan yang berkaitan dengan kesedihan, sakit hati, dan rasa prihatin terhadap peristiwa kebakaran hutan. Berdasarkan kedekatan maknanya, keempat topik tersebut dapat ditafsirkan sebagai satu tema yang lebih luas, yaitu "Respons Empati dan Kesedihan".

G. Hierarchical Clustering



Gambar 9. Dendrogram Hierarchical Clustering

Dendrogram pada Gambar 9 memperlihatkan pola pengelompokan yang sejalan dengan hasil Similarity Matrix. Pada jarak sekitar 0,5–0,6, Topik 6, 7, 11, dan 12 telah bergabung dalam satu klaster karena memiliki kemiripan makna yang tinggi. Pada jarak sekitar 0,7–0,8, topik yang berkaitan dengan doa, yaitu Topik 1 dan 5, mulai bergabung. Selanjutnya, pada jarak sekitar 1,0, sebagian besar topik emosional membentuk kelompok yang lebih besar, sedangkan Topik 13 dan Topik 8 masih tampak sebagai cabang yang terpisah. Hasil ini mengindikasikan bahwa 14 topik awal dapat diringkas secara hierarkis menjadi sekitar tiga hingga empat tema besar yang mendominasi percakapan mengenai Karhutla.

Temuan tersebut dapat dibandingkan dengan penelitian El-Gayar et al. (2024) yang membandingkan LDA, BERTopic, dan RoBERTa pada ulasan aplikasi kesehatan. Penelitian tersebut melaporkan bahwa BERTopic dan RoBERTa menghasilkan topik yang lebih kohesif dibandingkan LDA, serta kata-kata yang lebih terfokus sehingga proses pemberian label topik menjadi lebih mudah. Temuan itu mendukung penggunaan BERTopic dalam penelitian ini karena kata kunci dan kedekatan semantik yang dihasilkan membantu peneliti mengidentifikasi kelompok pembahasan dalam komentar Instagram. Meskipun demikian, hasil El-Gayar et al. (2024) berasal dari konteks ulasan aplikasi kesehatan, sehingga digunakan sebagai pembanding metodologis dan bukan sebagai bukti langsung mengenai isu Karhutla.

Secara keseluruhan, hasil pemodelan menunjukkan bahwa percakapan masyarakat di Instagram mengenai Karhutla di Indonesia, khususnya pada komentar yang membahas wilayah Kalimantan, tidak hanya berisi pembahasan mengenai kebakaran dan pihak yang dianggap bertanggung jawab, tetapi juga memuat doa, harapan, kritik, serta respons empati dan kesedihan. Penggabungan hasil distribusi topik, kata kunci c-TF-IDF, matriks kemiripan, dan dendrogram memberikan gambaran yang lebih menyeluruh mengenai tema-tema utama dalam percakapan publik.

4. Kesimpulan

Berdasarkan hasil penelitian, metode BERTopic mampu mengidentifikasi dan mengelompokkan percakapan masyarakat mengenai Karhutla berdasarkan kemiripan makna komentar Instagram. Dari 2.061 komentar yang dianalisis, diperoleh 14 topik valid dengan 1.808 komentar terkelompok dan 253 komentar sebagai *outlier*. Topik yang dominan berkaitan dengan kebakaran hutan, hewan, dan dugaan tindakan pembakaran yang

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

disengaja. Selain itu, percakapan masyarakat mencakup harapan turunnya hujan, kritik terhadap pemerintah, serta respons empati dan kesedihan terhadap dampak kebakaran. Hasil *similarity matrix* dan *hierarchical clustering* menunjukkan adanya hubungan semantik antara beberapa topik yang dapat dirangkum menjadi tema pembahasan lebih luas. Secara keseluruhan, BERTopic dapat dimanfaatkan untuk memetakan kecenderungan pembahasan mengenai Karhutla berdasarkan data komentar Instagram. Namun, penelitian ini memiliki keterbatasan karena data hanya berasal dari Instagram sehingga belum tentu mewakili pandangan masyarakat secara keseluruhan. Selain itu, hasil pengelompokan topik dapat dipengaruhi oleh pemilihan model *embedding* dan pengaturan parameter BERTopic. Oleh karena itu, hasil penelitian perlu dipahami sesuai dengan cakupan data dan konfigurasi model yang digunakan.

Reference

1. Anugrah, A. F., & Agatha, R. D. (2025). Utilizing IndoBERT and BERTopic to explore public opinion on BPS Instagram posts. *Journal of Applied Informatics and Computing*, *9*(5), 2836–2844. https://doi.org/10.30871/jaic.v9i5.10327
2. Ardhika, R. D., Santoso, A. D., & Elfandari, S. (2025). Social media for wildlife conservation campaign in South Sumatra, Indonesia. *Jurnal Kawistara*, *15*(1), 66–81. https://doi.org/10.22146/kawistara.87795
3. Arisman. (2020). Analisis tren kebakaran hutan dan lahan di Indonesia periode tahun 2015–2019. *Jurnal Sains Teknologi & Lingkungan*, *6*(1). https://doi.org/10.29303/jstl.v6i1.131
4. El-Gayar, O., Al-Ramahi, M., Wahbeh, A., Nasralah, T., & Elnoshokaty, A. (2024). A comparative analysis of the interpretability of LDA and LLM for topic modeling: The case of healthcare apps. *AMCIS 2024 Proceedings*. https://aisel.aisnet.org/amcis2024/health_it/health_it/22
5. Gökçimen, T., & Daş, B. (2024). Exploring climate change discourse on social media and blogs using a topic modeling analysis. *Heliyon*, *10*(11), e32464. https://doi.org/10.1016/j.heliyon.2024.e32464
6. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. https://doi.org/10.48550/arXiv.2203.05794
7. Hamdi, S., Rizal, S., Shibata, T., Darmawan, A., Irfan, M., & Sulaiman, A. (2025). The dispersion of smoke haze from peatland fires over South Sumatra during the moderate El Niño of 2023. *Natural Hazards*, *121*, 1095–1116. https://doi.org/10.1007/s11069-024-06857-x
8. Khairani, M. M., Purba, I. A., & Ginting, R. (2026). Peran media sosial Instagram @BPBDPROVINSISUMUT dan @INFOBMKGSUMUT dalam penyebaran informasi bencana banjir di Sumatera Utara. *Linimasa: Jurnal Ilmu Komunikasi*, *9*(1).
9. Lie, S. T., Rahman, S. D. F., Rachmad, A. D., Maruddani, D. A. I., & Kartikasari, P. (2026). Topic modeling and text classification of public comments towards BPJS Kesehatan using BERTopic and IndoBERT. *AIP Conference Proceedings*, *3411*(1), 040014. https://doi.org/10.1063/5.0322537
10. Mariwy, C. F., Watratan, C. J. G., Dos Santos, R., Reba, D. N., Sabono, R. F., & Pasiakan, N. I. (2025). Analisis sentimen komentar TikTok tentang kasus TPPO TKI di Kamboja menggunakan klasifikasi Naive Bayes untuk mengukur opini publik. *Techno.COM*, *24*(2), 602–616. https://doi.org/10.62411/tc.v24i2.12793
11. Mifrah, S., & Benlahmar, E. H. (2022). Topic modeling with transformers for sentence-level using coronavirus corpus. *International Journal of Interactive Mobile Technologies*, *16*(17), 50–59. https://doi.org/10.3991/ijim.v16i17.33281
12. Murthy, D., Kurz, S. E., Anand, T., Hornick, S., Lakuduva, N., & Sun, J. (2025). Examining hurricane-related social media topics longitudinally and at scale: A transformer-based approach. *PLOS ONE*, *20*(1), e0316852. https://doi.org/10.1371/journal.pone.0316852
13. Nurhayati, A. D., Saharjo, B. H., Sundawati, L., Syartinilia, & Cochrane, M. A. (2021). Forest and peatland fire dynamics in South Sumatra Province. *Forest and Society*, *5*(2), 591–603. https://doi.org/10.24259/fs.v5i2.14435
14. Sze, J. S., Jefferson, & Lee, J. S. H. (2019). Evaluating the social and environmental factors behind the 2015 extreme fire event in Sumatra, Indonesia. *Environmental Research Letters*, *14*, 015001. https://doi.org/10.1088/1748-9326/aaee1d
15. Uthirapathy, S., & Sandanam, D. (2023). Topic modelling and opinion analysis on climate change Twitter data using LDA and BERT model. *Procedia Computer Science*, *218*, 908–917. https://doi.org/10.1016/j.procs.2023.01.071
16. Verbeij, T., Beyens, I., Trilling, D., & Valkenburg, P. M. (2024). Happiness and sadness in adolescents' Instagram direct messaging: A neural topic modeling approach. *Social Media + Society*. https://doi.org/10.1177/20563051241229655
17. Yao, Y., & Pearce, W. (2025). The visual vernacular of climate change on Instagram: How modal convergence between image and text is changing the representation of climate solutions. *International Journal of Communication*, *19*, 2221–2246. https://ijoc.org/index.php/ijoc/article/view/24449
18. Zankadi, H., Idrissi, A., Daoudi, N., & Hilal, I. (2022). Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-11373-1
19. Zaskia, A., Saharjo, B. H., & Albar, I. (2023). Pengaruh curah hujan terhadap kejadian kebakaran hutan dan lahan di Kabupaten Banyuasin Provinsi Sumatera Selatan. *Jurnal Silvikultur Tropika*, *14*(3), 195–200. https://doi.org/10.29244/j-siltrop.14.03.195-200

DOI: https://doi.org/10.69693/ijmst.v5i1.14642
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)