PEMODELAN TOPIK CUITAN TENTANG DANANTARA MENGGUNAKAN BERTOPIC YANG DIOPTIMASI DENGAN OPTUNA



Disusun Oleh:

Nama : Noor Akhnafal Aban
NIM : 22523179

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
2026

HALAMAN PENGESAHAN DOSEN PEMBIMBING

PEMODELAN TOPIK CUITAN TENTANG DANANTARA MENGGUNAKAN BERTOPIC TEROPTIMASI UMAP DAN HDBSCAN

TUGAS AKHIR

Disusun Oleh:

N a m a	: Noor Akhnafal Aban
NIM	: 22523179

Yogyakarta, 13 Januari 2026

Pembimbing,

(Chanifah Indah Ratnasari, S.Kom., M.Kom.)

HALAMAN PENGESAHAN DOSEN PENGUJI

PEMODELAN TOPIK CUITAN TENTANG DANANTARA MENGGUNAKAN BERTOPIC TEROPTIMASI UMAP DAN HDBSCAN

TUGAS AKHIR

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer dari Program Studi Informatika – Program Sarjana di Fakultas Teknologi Industri Universitas Islam Indonesia

Yogyakarta, 13 Januari 2026

Tim Penguji

Chanifah Indah Ratnasari, S.Kom., M.Kom. _____

Anggota 1

Mukhammad Andri Setiawan, S.T., M.Sc., Ph.D. _____

Anggota 2

Arrie Kurniawardhani, S.Si., M.Kom. _____

Mengetahui,

Ketua Program Studi Informatika – Program Sarjana

Fakultas Teknologi Industri

Universitas Islam Indonesia

(Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Noor Akhnafal Aban

NIM : 22523179

Tugas akhir dengan judul:

PEMODELAN TOPIK CUITAN TENTANG DANANTARA MENGGUNAKAN BERTOPIC TEROPTIMASI UMAP DAN HDBSCAN

Menyatakan bahwa seluruh komponen dan isi dalam tugas akhir ini adalah hasil karya saya sendiri. Apabila di kemudian hari terbukti ada beberapa bagian dari karya ini adalah bukan hasil karya sendiri, tugas akhir yang diajukan sebagai hasil karya sendiri ini siap ditarik kembali dan siap menanggung risiko dan konsekuensi apapun.

Demikian surat pernyataan ini dibuat, semoga dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 13 Januari 2026



Noor Akhnafal Aban

HALAMAN PERSEMBAHAN

Karya ini saya persembahkan kepada kedua orang tua saya sebagai wujud penghargaan atas dukungan, doa, serta keteguhan mereka dalam membimbing setiap langkah yang saya tempuh. Kepada para dosen dan pembimbing yang telah memberikan arahan kritis, ruang untuk berkembang, serta standar akademik yang menuntut saya untuk berpikir lebih sistematis dan objektif.

Saya persembahkan pula untuk rekan-rekan yang menjadi mitra diskusi selama proses penelitian, yang membantu mempertajam analisis dan memastikan setiap keputusan metodologis, mulai dari pemilihan model representasi, pengurangan *noise* dalam korpus, hingga optimasi klastering.

Terakhir, karya ini saya dedikasikan untuk diri saya sendiri, sebagai pengingat atas konsistensi dalam mempelajari pemodelan topik, eksplorasi optimasi UMAP dan HDBSCAN, dan upaya memahami struktur informasi dalam data teks secara ilmiah. Semoga penelitian ini menjadi langkah awal menuju kontribusi yang lebih matang di bidang pemrosesan bahasa alami, pembelajaran mesin, dan pengembangan ilmu data.

HALAMAN MOTO

Pemahaman yang kuat selalu berangkat dari keraguan yang terkelola. Setiap asumsi wajib diuji, setiap keputusan metodologis harus memiliki dasar yang dapat dipertanggungjawabkan.

KATA PENGANTAR

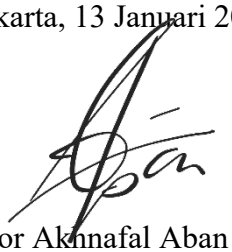
Puji syukur ke hadirat Allah Swt., karena atas rahmat dan karunia-Nya laporan tugas akhir ini dapat diselesaikan sesuai dengan tujuan penelitian. Penelitian ini mengkaji pemodelan topik pada data teks menggunakan pendekatan BERTopic dengan optimasi UMAP dan HDBSCAN, serta mengevaluasi pengaruh konfigurasi model terhadap kualitas struktur topik yang dihasilkan, guna memberikan pemahaman metodologis yang lebih sistematis mengenai ekstraksi topik berbasis representasi bahasa modern.

Selama pelaksanaan penelitian, terdapat beberapa kendala nonilmiah, terutama terkait penyesuaian aktivitas di luar akademik dan hambatan teknis terkait lingkungan sistem operasi. Namun, seluruh proses penelitian dapat diselesaikan dengan baik berkat dukungan dari pihak. Oleh karena itu, saya mengucapkan terima kasih kepada pihak yang secara langsung maupun tidak langsung membantu saya dalam menyelesaikan penelitian ini. Maka dari itu, mengucapkan terima kasih kepada:

1. Bapak Prof. Fathul Wahid, S.T., M.Sc., Ph.D. selaku Rektor Universitas Islam Indonesia atas dukungan dan kebijakan institusional dalam penyelenggaraan pendidikan.
2. Bapak Prof. Dr. Ir. Hari Purnomo, M.T., IPU, ASEAN.Eng selaku Dekan Fakultas Teknologi Industri atas dukungan akademik dan fasilitasi selama masa studi.
3. Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D. selaku Kepala Program Studi Informatika atas arahan dan dukungan dalam pelaksanaan proses akademik.
4. Ibu Chanifah Indah Ratnasari, S.Kom., M.Kom. selaku dosen pembimbing atas arahan kritis dan pendampingan selama proses bimbingan. Penulis juga menyampaikan permohonan maaf atas sikap dan komunikasi yang belum sepenuhnya mencerminkan etika akademik, serta menjadikan teguran yang diberikan sebagai pembelajaran penting.
5. Deva Soka Faridh Aziz dan Ainun Safitri atas kesediaan dan kontribusinya sebagai *human annotator* dalam memvalidasi hasil interpretasi topik.
6. Nurul Anwar dan Yuli Fitriani selaku orang tua penulis, serta keluarga yang senantiasa memberikan dukungan moral, doa, dan motivasi selama proses penyusunan laporan ini.
7. Teman-teman yang telah memberikan sarana berdiskusi teknis maupun nonteknis yang senantiasa menyempatkan waktunya untuk membantu saya selama menjalani penelitian.

Semoga penelitian ini dapat memberikan kontribusi bagi pengembangan studi pemodelan topik berbasis *transformer* dan menjadi referensi bagi penelitian lanjutan, khususnya pada analisis data teks berbahasa Indonesia yang tidak terstruktur dan memiliki tingkat *noise* tinggi.

Yogyakarta, 13 Januari 2026



Noor Akhnafal Aban

SARI

Penelitian ini bertujuan memetakan struktur wacana publik mengenai Danantara sebagai lembaga pengelola investasi negara melalui pemodelan topik berbasis BERTopic yang dioptimasi dengan UMAP dan HDBSCAN. Ketidakpastian informasi dan meningkatnya volume percakapan digital terkait Danantara menuntut pendekatan analitis yang mampu menangkap keragaman isu secara lebih komprehensif dibandingkan analisis sentimen konvensional. Data diperoleh dari himpunan cuitan yang telah melalui tahapan pra-pemrosesan, penghapusan duplikasi, serta deteksi anomali temporal untuk meminimalkan distorsi akibat ledakan percakapan tidak organik. BERTopic dikonfigurasi dengan *embedding* IndoSBERT-large yang secara spesifik dirancang untuk representasi semantik teks bahasa Indonesia, serta penyesuaian parameter UMAP–HDBSCAN guna menghasilkan struktur klaster yang stabil. Hasil pemodelan menghasilkan lima belas topik yang merepresentasikan spektrum wacana mulai dari dukungan terhadap penguatan ekonomi nasional, pembahasan energi dan BUMN, kritik terhadap isu legalitas dan akuntabilitas, hingga percakapan ringan yang bersifat humor. Temuan ini menunjukkan bahwa percakapan publik mengenai Danantara bersifat heterogen dan dipengaruhi oleh dinamika sosial-politik di media sosial. Secara keseluruhan, pendekatan BERTopic dengan optimalisasi UMAP–HDBSCAN menunjukkan efektivitas dalam mengungkap struktur diskursus yang berlapis berdasarkan tingkat koherensi dan keragaman topik yang dihasilkan serta menyediakan dasar berbasis data bagi analisis persepsi publik dan formulasi strategi komunikasi kebijakan.

Kata kunci: Danantara, pemodelan topik, BERTopic, UMAP, HDBSCAN.

GLOSARIUM

Topic modeling	Teknik analisis teks yang digunakan untuk mengidentifikasi dan mengekstraksi topik laten dari sekumpulan dokumen berdasarkan pola kemunculan kata, sehingga memungkinkan pemahaman struktur tematik dalam korpus teks secara otomatis.
Data scraping	Proses pengumpulan data secara otomatis dari sumber digital, seperti situs web atau platform media sosial, dengan mengekstraksi informasi yang relevan sesuai dengan kriteria tertentu untuk keperluan analisis.
Pra-pemrosesan	Tahapan awal dalam pengolahan data teks yang bertujuan untuk membersihkan dan menstandarkan data, meliputi langkah-langkah seperti normalisasi teks, penghapusan karakter tidak relevan, tokenisasi, serta penghilangan kata umum, guna meningkatkan kualitas analisis lanjutan.
Coherence	Metrik evaluasi yang digunakan dalam *topic modeling* untuk mengukur menunjukkan topik yang lebih koheren dan mudah diinterpretasikan secara konseptual.
Embedding	Representasi numerik berdimensi tetap yang memetakan unit teks ke dalam ruang vektor sehingga kemiripan semantik antar teks dapat diukur secara matematis.
UMAP	Metode reduksi dimensi nonlinear yang memproyeksikan data berdimensi tinggi ke ruang berdimensi lebih rendah dengan tetap mempertahankan struktur lokal dan global data.
HDBSCAN	Algoritma klasterisasi berbasis kepadatan yang mampu membentuk klaster dengan ukuran bervariasi serta mengidentifikasi data yang tidak termasuk ke dalam klaster sebagai *noise*.
Astroturfing	Aktivitas terkoordinasi yang dirancang untuk menciptakan kesan dukungan atau penolakan publik secara artifisial melalui distribusi pesan yang masif dan terstruktur.
Burst	Pola lonjakan kemunculan suatu token atau istilah dalam interval waktu sempit yang signifikan dibandingkan distribusi temporal normalnya, dan dalam penelitian ini digunakan sebagai indikator anomali temporal.

Hyperparameter	Parameter konfigurasi model yang ditetapkan sebelum proses pelatihan atau pemodelan dan tidak dipelajari secara langsung dari data.
Tuning	Proses penyesuaian nilai hyperparameter secara sistematis untuk memperoleh konfigurasi model yang menghasilkan kinerja optimal sesuai kriteria evaluasi.
Optuna	Kerangka kerja optimasi hyperparameter berbasis pencarian otomatis yang menggunakan pendekatan trial untuk mengevaluasi dan memilih konfigurasi parameter terbaik.

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN DOSEN PEMBIMBING ii
HALAMAN PENGESAHAN DOSEN PENGUJI iii
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR iv
HALAMAN PERSEMBAHAN v
HALAMAN MOTO vi
KATA PENGANTAR vii
SARI ix
GLOSARIUM x
DAFTAR ISI xii
DAFTAR TABEL xiv
DAFTAR GAMBAR xv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 4
1.3 Batasan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 5
1.6 Sistematika Penulisan 5
BAB II KAJIAN PUSTAKA 7
2.1 Danantara 7
2.2 Media Sosial X 7
2.2.1 Karakteristik Data Media Sosial X 8
2.2.2 Fenomena *Burst* 8
2.2.3 Pola Koordinasi dan Distorsi Wacana 8
2.3 Vektorisasi Teks 9
2.4 Topic Modeling 9
2.5 Perbandingan Pemodelan Topik 10
2.6 BERTopic 11
2.6.1 *Embedding* 12
2.6.2 Reduksi Dimensi 13
2.6.3 Klastering 13
2.6.4 *Class-Based* TF-IDF 14
2.7 Optimasi Pemodelan Topik 15
2.7.1 Sensitivitas Parameter UMAP 15
2.7.2 Sensitivitas Parameter HDBSCAN 15
2.7.3 Interaksi Kompleks antar Parameter 16
2.7.4 Pendekatan Optimasi 16
2.8 Penelitian Terdahulu 17
BAB III METODOLOGI PENELITIAN 19
3.1 Desain Metodologis dan Reprodusibilitas Penelitian 19
3.2 Pendefinisian Kebutuhan dan Analisis Awal 20
3.3 Proses Pengumpulan dan Deskripsi Hasil Data 22
3.4 Pemrosesan dan Pemodelan Topik Berbasis BERTopic 23
3.4.1 Pra-pemrosesan Teks 24
3.4.2 Pembersihan Data dan Reduksi *Noise* 24
3.4.3 Pembentukan *Embedding* dengan IndoSBERT 28

3.4.4 Reduksi Dimensi Menggunakan UMAP....28
3.4.5 Klasterisasi Dokumen Menggunakan HDBSCAN....29
3.4.6 Pembentukan Topik Menggunakan BERTopic melalui c-TF-IDF....29
3.5 Optimasi Parameter UMAP dan HDBSCAN melalui *Hyperparameter Tuning* .30
3.5.1 Ruang Pencarian Parameter....30
3.5.2 Kerangka Hyperparameter Tuning Berbasis Optuna....33
3.5.3 Perumusan Fungsi Objektif....33
3.6 Finalisasi Model....34
BAB IV HASIL DAN PEMBAHASAN....35
4.1 Ruang Lingkup Pembahasan....37
4.2 Karakteristik *Dataset* Hasil Pengumpulan....37
4.3 Pembersihan Data dan Deteksi Anomali Temporal....40
4.3.1 Penghapusan Duplikasi dan Spam Data....41
4.3.2 Deteksi dan Reduksi *Burst* Temporal....42
4.3.3 *Dataset* Pasca Pembersihan Temporal....47
4.3.4 Penyempurnaan Teks melalui Penghapusan Stopwords....47
4.4 Representasi Semantik Dokumen....48
4.5 *Hyperparameter tuning* BERTopic....49
4.5.1 Percobaan 1 – Optimasi Berbasis *Coherence*....49
4.5.2 Percobaan 2 – Penambahan Penalti Jumlah Topik....50
4.5.3 Percobaan 3 – Kombinasi *Coherence* dan *Diversity*....53
4.5.4 Percobaan 4 – Final Konfigurasi....56
4.5.5 Perbandingan Hasil Optuna....57
4.6 Struktur Topik Akhir Hasil Pemodelan....58
4.6.1 Distribusi Topik dan Ukuran Klaster....59
4.6.2 Representasi Kata Kunci Berbobot Topik....60
4.6.3 *Intertopic Distance Map*....61
4.7 Interpretasi Topik....62
4.7.1 Klaster *Outlier* dan Konten NonRelevan....69
4.7.2 Narasi Dukungan dan Optimisme Ekonomi....70
4.7.3 Kekhawatiran Terhadap Tata Kelola dan Korupsi....70
4.7.4 Dinamika Politik dan Legitimasi Kekuasaan....71
4.7.5 Isu Teknis Fiskal dan Investasi....72
4.7.6 Kebingungan Publik dan Percakapan Informal....73
4.8 Diskusi dan Implikasi....74
4.9 Kelebihan dan Kekurangan Penelitian....76
BAB V KESIMPULAN DAN SARAN....78
5.1 Kesimpulan....78
5.2 Saran....79
DAFTAR PUSTAKA....80
LAMPIRAN....86

DAFTAR TABEL

Tabel 2.1 Bingkai Analisis Pemodelan Topik 17
Tabel 4.1 Distribusi Frekuensi Token 39
Tabel 4.2 Perbandingan Sebelum dan Sesudah Pra-pemrosesan Data 41
Tabel 4.3 Ringkasan *Burst Cluster* 43
Tabel 4.4 Metrik Temporal Term Abnormal 44
Tabel 4.5 Ringkasan Term yang Dihapus 45
Tabel 4.6 Ringkasan Hasil Tuning Berbasis *Coherence* 50
Tabel 4.7 Ringkasan Hasil Tuning dengan Penalti Jumlah Topik 52
Tabel 4.8 Kombinasi *coherence, diversity, dan penalty* 54
Tabel 4.9 Hasil Optuna *Trials* Percobaan Ketiga 55
Tabel 4.10 Final konfigurasi 56
Tabel 4.11 Perbandingan Hasil Optuna Tiap Percobaan 57
Tabel 4.12 Dokumen Representatif 63
Tabel 4.13 Interpretasi Topik Mandiri 66
Tabel 4.14 Interpretasi Pasca Validasi dari *Annotator* 68

DAFTAR GAMBAR

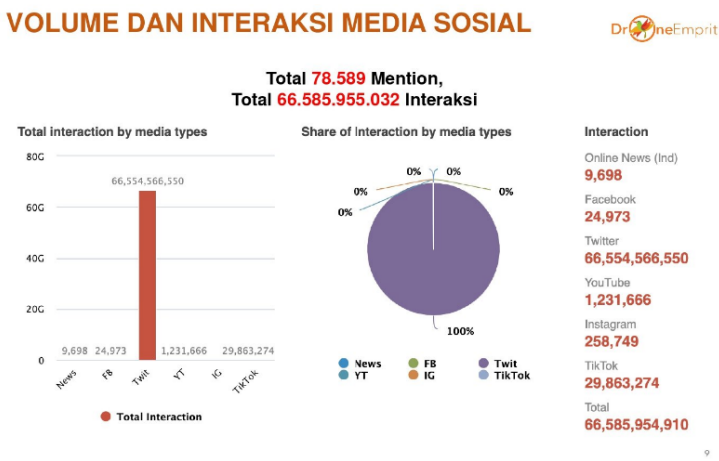
Gambar 1.1 Volume dan Interaksi Media Sosial 1
Gambar 2.1 Bagaimana BERTopic Bekerja 12
Gambar 3.1 Diagram Alir Penelitian 20
Gambar 3.2 *Workflow* n8n 22
Gambar 3.3 Hasil Pemodelan Topik Awal 24
Gambar 3.4 Kode Pembentukan Embedding 28
Gambar 3.5 Diagram Alir *Hyperparameter Tuning* dengan Optuna 30
Gambar 4.1 Statistik Deskriptif *Dataset* Mentah 37
Gambar 4.2 Distribusi Waktu Publikasi Cuitan 38
Gambar 4.3 Kepadatan kemunculan Konten 38
Gambar 4.4 Isi Kode pra-pemrosesan 41
Gambar 4.5 Statistik *Dataset* Setelah Deduplikasi 41
Gambar 4.6 Statistik *Dataset* Setelah Pembersihan 42
Gambar 4.7 Statistik Pasca Reduksi Noise 46
Gambar 4.8 Statistik *Dataset* Setelah Penghapusan *Stopwords* 47
Gambar 4.9 Statistik *Dataset* Akhir Pasca Deduplikasi Ulang 48
Gambar 4.10 Pratinjau Hasil *Embedding* dan Teks Terkait 48
Gambar 4.11 Kode Perumusan Penalti Jumlah Topik 51
Gambar 4.12 Distribusi Dokumen per Topik Final 59
Gambar 4.13 *Topic Word Scores Bar Chart* 60
Gambar 4.14 *Intertopic Distance Map* 61

BAB I
PENDAHULUAN

1.1 Latar Belakang

Danantara atau Daya Anagata Nusantara merupakan badan pengelola investasi yang dibentuk oleh Pemerintah Republik Indonesia dan secara resmi diluncurkan pada 24 al-ha Februari 2025. Lembaga ini dirancang untuk mengonsolidasikan dan mengoptimalkan aset negara guna mendukung pembiayaan proyek strategis nasional serta memperkuat pertumbuhan ekonomi jangka panjang (Badan Pengelola Investasi Daya Anagata Nusantara (BPI Danantara), 2025; Kementerian Keuangan Republik Indonesia, 2025). Pembentukan Danantara ditujukan sebagai instrumen keuangan negara untuk meningkatkan efisiensi pengelolaan kekayaan negara, memperluas kapasitas investasi, serta menciptakan pertumbuhan ekonomi (Mahrunisa dkk., 2025). Namun, sebagai institusi baru dengan pengelolaan aset yang sangat besar, Danantara menghadapi tantangan legitimasi publik yang berkaitan dengan isu transparansi, akuntabilitas, dan tata kelola (Priatna dkk., 2025).

Sejak pembentukannya, Danantara menjadi topik yang ramai diperbincangkan di berbagai platform media sosial, khususnya pada platform media sosial X. (Fahmi, 2025). Hal ini ditunjukkan pada Gambar 1.1 yang menunjukkan bahwa total interaksi publik terhadap Danantara terkonsentrasi secara sangat dominan pada platform Twitter atau sekarang disebut X dengan lebih dari 66,5 miliar interaksi, jauh melampaui kontribusi platform media sosial lainnya.



Gambar 1.1 Volume dan Interaksi Media Sosial

Sumber: Fahmi, 2025

Dominasi interaksi yang sangat besar pada platform X menunjukkan bahwa media sosial ini berperan sebagai media utama pembentukan dan pertukaran opini publik terkait Danantara, sekaligus menyediakan volume percakapan yang representatif untuk dianalisis. Tidak hanya pembahasan mengenai Danantara, media sosial X juga umum digunakan sebagai ruang publik virtual untuk percakapan masyarakat mengenai isu politik, kebijakan, dan isu nasional secara *real-time* (Ihsan & Adnan, 2022).

Kondisi tersebut menjadikan dinamika percakapan digital terutama dalam media sosial X menjadi aspek yang krusial untuk dianalisis lebih lanjut. Persepsi masyarakat terhadap kebijakan dan institusi negara di media sosial sangat dipengaruhi oleh arus informasi yang cepat, terbuka, dan rentan terhadap disinformasi, yang pada akhirnya berpotensi memengaruhi tingkat kepercayaan publik terhadap pemerintah dan lembaga publik (Rahmanto, 2022; Suriadi, 2025). Persepsi terhadap transparansi anggaran dan akuntabilitas pengelolaan dana negara, misalnya, merupakan faktor penting dalam membangun dan mempertahankan legitimasi lembaga pemerintah (Zein dkk., 2025). Dalam ruang publik virtual, persepsi masyarakat terhadap kebijakan dan institusi negara terartikulasikan melalui wacana yang terbentuk dari beragam isu, narasi, dan *framing* yang beredar (Yunanto dkk., 2024). Situasi ini menunjukkan bahwa kompleksitas wacana publik di media sosial menuntut pendekatan analitis yang mampu mengungkap struktur isu secara menyeluruh. Oleh karena itu, pemetaan wacana publik terhadap Danantara dari sumber data media sosial X berupa cuitan menjadi langkah yang relevan untuk memahami isu apa saja yang dibicarakan dan bagaimana isu tersebut terstruktur.

Sejumlah penelitian terdahulu mengenai Danantara menunjukkan bahwa kajian akademik yang berkembang sejauh ini masih didominasi oleh pendekatan analisis sentimen dalam memetakan persepsi publik. Beberapa studi, antara lain yang dilakukan oleh Njurumana dkk. (2025) memanfaatkan algoritma *Support Vector Machine* untuk mengukur persepsi publik terhadap Danantara, sementara penelitian oleh Hermanto dkk. (2025) mengidentifikasi dinamika opini publik yang beragam melalui pendekatan sentimen analisis dengan pendekatan yang sama. Pendekatan ini umumnya memanfaatkan algoritma klasifikasi teks untuk mengelompokkan opini ke dalam kategori polaritas seperti positif, negatif, atau netral, sehingga mampu memberikan gambaran mengenai kecenderungan sikap masyarakat terhadap pembentukan dan pengelolaan Danantara.

Namun demikian, hasil analisis sentimen bertumpu pada polaritas emosional yang membatasi ruang analisis untuk menangkap keragaman isu yang muncul dalam diskursus

publik (Yunanto dkk., 2024). Pendekatan ini cenderung mereduksi opini publik ke dalam kategori sederhana tanpa mampu menangkap relasi semantik antara ekspresi evaluatif dan isu yang dibicarakan (Veantiny dkk., 2023). Dalam konteks Danantara, satu percakapan dapat memuat beragam narasi sekaligus, seperti dukungan terhadap potensi ekonomi nasional dan kritik terhadap tata kelola atau aktor politik tertentu, yang tidak selalu sejalan secara emosional. Narasi yang beragam ini masuk ke dalam satu label sentimen, sehingga berpotensi meningkatkan *noise* dan menghasilkan kesimpulan yang kurang akurat (Bestvater & Monroe, 2023; Yunanto dkk., 2024). Dalam kondisi tersebut, analisis sentimen tidak mampu mengungkap substansi isu yang menjadi objek perbincangan publik karena hanya menjawab sikap atau polaritas emosional, bukan mengidentifikasi isu tematik yang membentuk sikap tersebut. Oleh karena itu, diperlukan pendekatan yang secara eksplisit berfokus pada identifikasi dan pemetaan isu tematik dalam diskursus publik.

Untuk melengkapi keterbatasan analisis sentimen yang berfokus pada polaritas emosional, pendekatan pemodelan topik modern menawarkan kerangka analisis yang berorientasi pada identifikasi struktur tematik dalam korpus teks pendek. BERTopic mengombinasikan *embedding* berbasis *transformer* dengan teknik klasterisasi HDBSCAN untuk membentuk topik yang relatif stabil dan koheren dan juga representasi topik melalui c-TF-IDF (Grootendorst, 2022). Studi komparatif menunjukkan bahwa BERTopic memiliki kinerja yang lebih baik dibandingkan metode pemodelan topik lain dalam konteks data media sosial X (Egger & Yu, 2022). Penelitian lain juga menekankan pentingnya optimasi parameter klasterisasi berbasis *coherence-aware optimization* untuk meningkatkan kualitas dan interpretabilitas topik yang dihasilkan (Syahrial & Afidh, 2024). Selain itu, kajian mengenai representasi semantik menegaskan bahwa *embedding* kontekstual berbasis *transformer* memberikan keunggulan dalam menangkap makna dan nuansa wacana digital (Devlin dkk., 2019).

Meskipun demikian, sejauh penelusuran literatur yang dilakukan, belum ditemukan penelitian yang secara khusus memodelkan topik percakapan publik mengenai Danantara menggunakan BERTopic. Padahal, analisis topik memiliki potensi untuk memetakan isu tematik dan distribusi wacana publik yang relevan. Ketiadaan kajian tersebut menunjukkan adanya celah penelitian, khususnya dalam upaya memahami wacana publik mengenai Danantara secara lebih komprehensif.

Berdasarkan pertimbangan tersebut, penelitian ini diposisikan untuk melakukan pemodelan topik terhadap percakapan publik tentang Danantara di platform media sosial X

menggunakan BERTopic yang dioptimasi melalui UMAP dan HDBSCAN. Penelitian ini bertujuan untuk mengidentifikasi topik-topik utama yang muncul dan menganalisis wacana publik yang terbentuk terhadap Danantara. Dengan pendekatan kuantitatif berbasis analisis teks, penelitian ini tidak dimaksudkan untuk menilai kebenaran atau dampak kebijakan, melainkan untuk memetakan struktur diskursus publik sebagai dasar analitis bagi pemahaman kecenderungan persepsi masyarakat secara tidak langsung.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

a. Bagaimana karakteristik data percakapan publik mengenai Danantara di platform media sosial X?
b. Bagaimana kualitas pemodelan topik yang dihasilkan melalui optimasi UMAP dan HDBSCAN pada model BERTopic?
c. Topik-topik apa saja yang dihasilkan dan isu-isu apa yang direpresentasikan oleh masing-masing topik serta bagaimana distribusi wacana publik terhadap Danantara tercermin melalui keterkaitan antartopik tersebut?

1.3 Batasan Masalah

Untuk menjaga fokus dan keterkendalian penelitian, batasan masalah yang diterapkan adalah sebagai berikut:

a. Data penelitian dibatasi pada percakapan publik di platform media sosial X yang diperoleh melalui proses *scraping*.
b. Pembersihan data dilakukan melalui penghapusan duplikasi dan reduksi *noise* berbasis anomali temporal, tanpa melakukan analisis lanjutan jaringan koordinasi antarakun.
c. Analisis difokuskan pada pemodelan topik menggunakan BERTopic dan tidak mencakup analisis sentimen, deteksi sikap, maupun inferensi kausal.
d. Interpretasi hasil penelitian bersifat deskriptif dan bertujuan memetakan struktur wacana, bukan menilai kebenaran isi percakapan.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

a. Menganalisis karakteristik data percakapan publik mengenai Danantara di platform media sosial X.
b. Memperoleh konfigurasi pemodelan topik yang optimal melalui optimasi parameter UMAP dan HDBSCAN pada model BERTopic menggunakan *Hyperparameter tuning* dengan Optuna.
c. Mengidentifikasi topik-topik yang dihasilkan dari pemodelan BERTopic, menafsirkan isu-isu tematik yang direpresentasikan oleh masing-masing topik, dan menganalisis distribusi wacana publik mengenai Danantara berdasarkan keterkaitan antartopik.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini meliputi:

a. Manfaat teoretis, yaitu memberikan kontribusi pada kajian pemodelan topik dan analisis wacana digital berbahasa Indonesia, khususnya dalam konteks kebijakan publik.
b. Manfaat praktis, yaitu menyediakan pemetaan isu dan struktur wacana publik yang dapat dimanfaatkan oleh pemangku kebijakan, peneliti, dan praktisi komunikasi publik sebagai bahan pertimbangan dalam merumuskan strategi komunikasi dan kebijakan yang lebih responsif terhadap opini masyarakat, dalam hal ini mengenai Danantara.

1.6 Sistematika Penulisan

Sistematika penulisan tugas akhir ini disusun untuk memberikan gambaran runtut mengenai alur penelitian, mulai dari perumusan masalah hingga penarikan kesimpulan. Penulisan laporan ini terbagi ke dalam lima bab sebagai berikut.

BAB I PENDAHULUAN

Bab ini memuat latar belakang penelitian yang menjelaskan konteks pembentukan Danantara dan dinamika percakapan publik di media sosial X, rumusan masalah yang menjadi fokus penelitian, tujuan penelitian, batasan penelitian, manfaat penelitian, serta sistematika penulisan.

BAB II KAJIAN PUSTAKA

Bab ini berisi landasan teori dan tinjauan pustaka yang relevan dengan penelitian, meliputi konsep wacana publik di media sosial, pemodelan topik, pendekatan *neural topic modeling*, BERTopic, UMAP, HDBSCAN, serta penelitian-penelitian terdahulu yang berkaitan dengan analisis opini publik dan pemodelan topik berbasis teks.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian yang digunakan secara sistematis, mencakup sumber dan teknik pengumpulan data, tahapan pra-pemrosesan teks, pembersihan data dan reduksi *noise* termasuk deteksi anomali temporal, pembentukan *embedding* menggunakan IndoSBERT-large, perumusan fungsi objektif, proses *hyperparameter tuning* menggunakan Optuna, serta tahapan pemodelan topik dengan BERTopic.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian yang diperoleh dari seluruh tahapan analisis, meliputi deskripsi data, hasil pra-pemrosesan dan pembersihan data, hasil deteksi anomali temporal, hasil pemodelan topik menggunakan *hyperparameter tuning*, serta visualisasi dan interpretasi struktur topik yang terbentuk. Pembahasan difokuskan pada analisis tematik wacana publik mengenai Danantara berdasarkan topik-topik yang dihasilkan.

BAB V KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan penelitian yang merangkum temuan utama sesuai dengan tujuan dan rumusan masalah, serta saran yang ditujukan bagi penelitian selanjutnya dan pihak-pihak terkait berdasarkan hasil analisis yang telah dilakukan.

BAB II
KAJIAN PUSTAKA

2.1 Danantara

Danantara sebagai badan pengelola investasi secara resmi diluncurkan pada 24 Februari 2025. Lembaga ini dirancang untuk mengonsolidasikan dan mengoptimalkan aset negara guna mendukung pembiayaan proyek strategis nasional serta memperkuat pertumbuhan ekonomi jangka panjang (Badan Pengelola Investasi Daya Anagata Nusantara (BPI Danantara), 2025; Kementerian Keuangan Republik Indonesia, 2025). Pembentukan Danantara ditujukan sebagai instrumen keuangan negara untuk meningkatkan efisiensi pengelolaan kekayaan negara, memperluas kapasitas investasi, serta menciptakan pertumbuhan ekonomi (Mahrunisa dkk., 2025).

2.2 Media Sosial X

Media sosial X telah menjadi ruang publik virtual yang memungkinkan masyarakat untuk terlibat secara langsung dalam percakapan mengenai isu politik, kebijakan, dan isu nasional secara terbuka dan *real-time* (Ihsan & Adnan, 2022). Dalam konteks Danantara, platform ini memiliki interaksi terbanyak dalam membahas Danantara dibandingkan dengan media sosial lain Fahmi (2025). Media sosial X memfasilitasi pertukaran opini yang cepat dan spontan sehingga menjadi sumber data yang relevan untuk mengamati dinamika persepsi publik. Diskursus yang terbentuk memperlihatkan bagaimana isu tertentu dibingkai, diperdebatkan, dan dikaitkan dengan isu lain. Oleh karena itu, analisis terhadap percakapan di media sosial X dapat memberikan gambaran yang lebih luas mengenai pola wacana publik.

Namun, cuitan dari media sosial X juga menghadirkan tantangan analitis. Cuitan bersifat pendek, tidak terstruktur, bersifat *noisy* dan sering kali mengandung bahasa informal, ironi, atau konteks implisit (Barreto dkk., 2021). Selain itu, diskursus di media sosial rentan terhadap dominasi isu tertentu, pengulangan konten, serta gangguan berupa spam dan aktivitas tidak organik (Schoch dkk., 2022; Singh & Kumari, 2021). Kondisi ini menuntut pendekatan analisis teks yang mampu menangkap keberagaman isu yang muncul secara bersamaan. Dengan mempertimbangkan karakteristik tersebut, media sosial X adalah sumber data yang relevan untuk mengkaji persepsi publik terhadap Danantara dan wacana yang terbentuk.

2.2.1 Karakteristik Data Media Sosial X

Kualitas data media sosial dapat memengaruhi proses pemodelan topik karena struktur topik yang dihasilkan pada dasarnya mengikuti pola pembentukan korpus sendiri (Churchill & Singh, 2022). Kualitas data tersebut tercermin pada karakteristik data media sosial itu sendiri. Karakteristik tersebut tampak dalam berbagai fenomena khas media sosial yang secara langsung membentuk dominasi topik dalam korpus. Fenomena tersebut seperti ledakan percakapan yang tidak natural serta pola aktivitas yang terkoordinasi (Schoch dkk., 2022; Singh & Kumari, 2021). Hal ini sering kali berkaitan dengan upaya pihak tertentu untuk memperkuat legitimasi politik dalam media sosial (Daeni dkk., 2023).

2.2.2 Fenomena *Burst*

Dalam konteks media sosial X, fenomena *burst* ini merupakan peningkatan mendadak volume percakapan terkait isu tertentu dalam rentang waktu yang singkat (Goswami dkk., 2024). Fenomena ini umumnya dipicu oleh isu yang ramai diperbincangkan dan dapat menyebabkan terbentuknya topik yang sangat temporal. Terdapat dugaan juga bahwa *burst* tercipta melalui oknum yang memanfaatkan botnet untuk membuat cuitan berulang yang serupa (Perdana dkk., 2015). Fenomena semacam ini akan mengakibatkan hasil pemodelan cenderung mendominasi terhadap satu topik. Kondisi tersebut berpotensi mengaburkan topik-topik lain yang lebih persisten dan substantif dalam wacana publik (Singh & Kumari, 2021). Oleh karena itu, penelitian ini perlu untuk mengidentifikasi *burst* yang terindikasi tidak organik, khususnya dugaan terkait dengan kepentingan politik, agar korpus yang digunakan lebih bersih dan struktur topik yang dihasilkan tetap stabil (Li Q dkk., 2023).

2.2.3 Pola Koordinasi dan Distorsi Wacana

Fenomena *burst* yang telah dibahas sebelumnya berkaitan erat dengan aktivitas terkoordinasi di media sosial, seperti kampanye informasi terstruktur atau *astroturfing* (García dkk., 2022). Dalam wacana Indonesia, pihak yang melakukan koordinasi seperti ini kerap dikenal dengan istilah *buzzer*. Aktivitas yang dilakukan *buzzer* diduga memicu lonjakan percakapan tidak natural dan membentuk klaster topik yang merefleksikan pola koordinasi, bukan opini publik yang organik (Schoch dkk., 2022).

Dalam penelitian ini, kajian mengenai pola koordinasi digunakan sebagai landasan konseptual untuk memahami sumber *burst* yang tidak organik. Namun, pendekatan yang diterapkan tetap difokuskan pada pembersihan *burst* percakapan, bukan pada identifikasi

jaringan *astroturfing* secara eksplisit, sehingga pemodelan topik diarahkan pada pembentukan struktur topik yang lebih stabil dan representatif.

2.3 Vektorisasi Teks

Sebelum dilakukan *Topic Modeling,* data perlu untuk dikonversi ke dalam bentuk representasi numerik yang dalam konteks ini adalah vektor. Vektorisasi adalah teknik untuk memetakan data ke dalam vektor berdimensi tertentu. Vektorisasi yang umum digunakan pada pemodelan klasik adalah *Bag-of-Words* (BOW) dan *Term Frequency–Inverse Document Frequency* (TF-IDF) (Janssens dkk., 2025). Meskipun efektif pada korpus tertentu, vektorisasi berbasis frekuensi memiliki keterbatasan mendasar ketika diterapkan pada data media sosial yang bersifat *noisy* seperti media sosial X. Representasi ini mengabaikan urutan kata dan struktur sintaks dan tidak menangkap relasi semantik antarkata (Yan dkk., 2020). Selain BOW dan TF-IDF, metode *embedding* dapat merepresentasikan teks secara kontekstual. Model *embedding* berbasis *transformer* seperti BERT membentuk representasi kata dan kalimat dengan mempertimbangkan konteks bidireksional, sehingga makna tidak direduksi menjadi frekuensi kemunculan semata (Devlin dkk., 2019).

2.4 Topic Modeling

Topic Modeling (Pemodelan topik) merupakan pendekatan analisis teks yang bertujuan mengidentifikasi topik tersembunyi dan memungkinkan pemetaan struktur semantik suatu korpus dokumen (Weston dkk., 2023). Metode pemodelan topik yang umum digunakan adalah sebagai berikut:

a. *Latent Dirichlet Allocation* (LDA)

   Latent Dirichlet Allocation merupakan model probabilistik generatif yang menggunakan dokumen sebagai campuran topik laten, di mana setiap topik direpresentasikan oleh distribusi probabilitas kata (Blei dkk., 2003). Proses generatif LDA mengasumsikan bahwa dokumen dihasilkan dengan terlebih dahulu memilih distribusi topik, kemudian untuk setiap kata memilih topik tertentu dan selanjutnya memilih kata berdasarkan distribusi topik tersebut. Model ini menggunakan asumsi bag-of-words, sehingga urutan kata dalam dokumen tidak diperhitungkan (Socs, 2018).

b. Top2Vec

   Top2Vec adalah metode pemodelan topik berbasis *embedding* yang secara otomatis mendeteksi topik dengan memetakan dokumen, kata, dan topik ke dalam ruang vektor

bersama, di mana kedekatan spasial merepresentasikan kemiripan semantik (Angelov, 2020). Metode ini membangun *embedding* dokumen dan kata menggunakan model representasi seperti Doc2Vec, *Universal Sentence Encoder*, atau BERT, kemudian memetakan dokumen ke dalam ruang vektor berdimensi tinggi. Topik diperoleh dengan mengidentifikasi area dokumen yang padat secara semantik menggunakan algoritma *klastering* HDBSCAN, di mana kata-kata yang berada dekat dengan pusat klaster digunakan sebagai representasi topik (Pangastuti dkk., 2025).

c. BERTopic

BERTopic merupakan metode pemodelan topik berbasis neural *(Neural topic modeling)* yang mengombinasikan *embedding transformer* seperti BERT dengan reduksi dimensi UMAP dan klastering HDBSCAN untuk menghasilkan topik yang koheren dan mudah diinterpretasikan (Grootendorst, 2022). Proses BERTopic dimulai dengan membentuk *embedding* dokumen menggunakan model *transformer*, kemudian mereduksi dimensi *embedding* tersebut menggunakan UMAP untuk mempertahankan struktur semantik global. Selanjutnya, dokumen dikelompokkan menggunakan HDBSCAN berdasarkan kemiripan semantik, dan representasi topik diekstraksi menggunakan *class-based* TF-IDF untuk menonjolkan kata yang paling khas pada setiap klaster (Wahyuni dkk., 2025).

2.5 Perbandingan Pemodelan Topik

Dalam konteks perbandingan, penelitian ini tidak melakukan perbandingan antar penelitian atau menilai metode berdasarkan rerata skor metrik seperti *coherence* atau *diversity* semata. Pendekatan seperti itu berisiko bias karena nilai metrik dapat berubah ketika perbandingan dilakukan pada *dataset* yang berbeda. Sebagai contoh, perbandingan yang disajikan oleh Grootendorst (2022) menunjukkan variasi nilai *topic coherence* dan *topic diversity* ketika model dievaluasi pada *dataset* 20 *Newsgroups* dan *BBC News* dengan metode pra-pemrosesan yang sama.

Dengan mempertimbangkan keterbatasan tersebut, penilaian metode pemodelan topik didasarkan pada temuan penelitian. Salah-satunya adalah studi komparatif oleh Egger & Yu (2022), yang mengevaluasi kinerja empat metode pemodelan yaitu *latent Dirichlet allocation* (LDA), *nonnegative matrix factorization* (NMF), Top2Vec, dan BERTopic. Berdasarkan Egger & Yu (2022), BERTopic dinilai sebagai metode pemodelan topik dengan performa paling kuat karena menunjukkan kinerja yang relatif seimbang pada berbagai aspek.

Keunggulan utama BERTopic terletak pada integrasi *embedding* kontekstual dengan mekanisme *class-based* TF-IDF (c-TF-IDF) yang mampu mengekstraksi kata representatif secara lebih diskriminatif dan bermakna dibandingkan pendekatan berbasis frekuensi murni (Kwon dkk., 2024). Dibandingkan dengan Top2Vec, BERTopic cenderung menghasilkan struktur topik yang lebih terpisah dan terdefinisi dengan baik, sementara Top2Vec sering menghasilkan topik yang saling tumpang tindih dan mencakup beberapa konsep sekaligus akibat ketergantungannya pada kepadatan spasial *embedding* tanpa mekanisme pembobotan kata yang eksplisit (Egger & Yu, 2022).

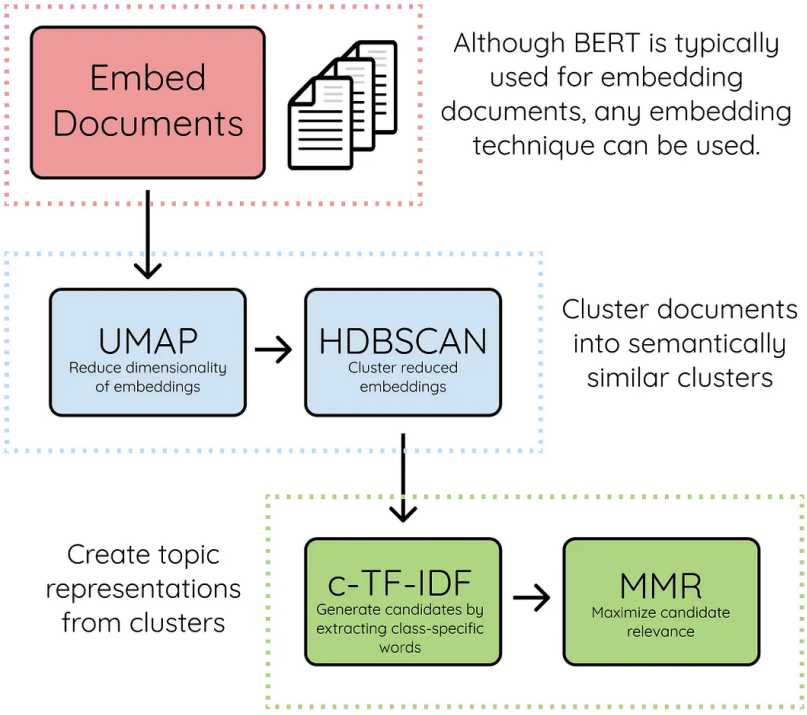
Sementara itu, LDA memiliki keterbatasan signifikan ketika diterapkan pada data pendek dan *noisy* (Albanese, 2022). LDA dikenal stabil dan interpretatif pada korpus dokumen panjang serta mampu merepresentasikan satu dokumen sebagai campuran beberapa topik, namun performanya menurun pada teks pendek dan sangat bergantung pada penentuan jumlah topik serta penyetelan hyperparameter yang ekstensif (Egger & Yu, 2022; Portuna, 2024; Socs, 2018). NMF, meskipun dalam beberapa studi dilaporkan memiliki kinerja yang lebih baik dibandingkan LDA dalam kondisi tertentu (Chen dkk., 2019), pada dasarnya merupakan metode faktorisasi matriks umum yang tidak dirancang secara inheren untuk pemodelan topik dan banyak digunakan untuk tujuan lain di luar analisis tematik (Guo dkk., 2024; Movahed dkk., 2026). Keunggulan LDA dan NMF dapat merepresentasikan dokumen ke dalam beberapa topik. Namun, hal tersebut tidak menjadi kebutuhan utama dalam penelitian ini, karena fokus analisis diarahkan pada pemetaan isu dan struktur wacana secara global.

Dengan mempertimbangkan keseimbangan antara kemampuan menangkap konteks semantik, kualitas representasi topik, serta kesesuaian dengan karakteristik data teks pendek dan heterogen, BERTopic dipandang sebagai metode pemodelan topik yang paling relevan untuk digunakan dalam penelitian ini, meskipun disertai konsekuensi berupa kompleksitas komputasi dan kebutuhan sumber daya yang lebih besar (Egger & Yu, 2022; Indriani, 2024).

2.6 BERTopic

BERTopic merupakan teknik pemodelan topik neural yang memanfaatkan perkembangan terkini dalam kecerdasan buatan untuk meningkatkan kualitas analisis topik. Dengan mengandalkan representasi semantik dari *embedding*, BERTopic mampu menghasilkan topik yang lebih koheren dan relatif mudah diinterpretasikan. Metode ini mengombinasikan *embedding* dengan pendekatan *class-based* TF-IDF untuk mengidentifikasi klaster dokumen yang padat secara semantik. Kepadatan klaster tersebut memungkinkan

pembentukan topik yang jelas dan representatif, sekaligus mempertahankan kata-kata yang paling relevan dalam menggambarkan setiap topik (Albanese, 2022). Proses pemodelan ini terdiri atas beberapa tahapan utama yang dirangkum pada Gambar 2.1 Bagaimana BERTopic Bekerja sebagai berikut.



Gambar 2.1 Bagaimana BERTopic Bekerja

Sumber: BERTopic - Advanced Transformer-Based Topic Modeling (t.t.)

2.6.1 *Embedding*

Embedding merupakan proses untuk mengonversi sebuah teks yang menjadi vector dengan menyimpan konteks semantik teks. Terdapat model *embedding* yang dilatih khusus pada korpus berbahasa Indonesia, yang dalam hal ini akan bermanfaat untuk merepresentasikan karakteristik kebahasaan lokal dibandingkan model yang dilatih menggunakan banyak bahasa (Wang dkk., 2024). IndoSBERT merupakan model *sentence embedding* yang dikembangkan dari IndoBERT dan dirancang untuk merepresentasikan kesamaan semantik antarkalimat melalui arsitektur *Siamese Networks* (Diana & Khodra, 2023). Studi yang dilakukan oleh Diana & Khodra (2023) menunjukkan bahwa IndoSBERT menghasilkan representasi kalimat berbahasa Indonesia yang lebih bermakna dan akurat dalam menangkap kesamaan semantik dibandingkan model IndoBERT sebelumnya maupun pendekatan *embedding* multibahasa, sehingga relevan untuk analisis teks pendek dan kontekstual. Dalam kerangka BERTopic,

embedding kontekstual seperti IndoSBERT diposisikan sebagai fondasi utama pembentukan topik sebelum dimensi reduksi dan klastering (Grootendorst, 2022).

2.6.2 Reduksi Dimensi

Pada tahap *embedding*, setiap kata atau kalimat diubah menjadi vektor dalam ruang berdimensi tinggi, dengan ukuran vektor yang umumnya mencapai ratusan dimensi. Dimensi yang sangat besar ini membuat data menjadi kompleks, terutama ketika jumlah kata atau kalimat dalam *dataset* sangat banyak. Akibatnya, beberapa algoritma klastering mengalami kesulitan dalam menemukan pola atau kesamaan antar data karena kompleksitas ruang vektor yang tinggi.

Untuk mengatasi hal tersebut, digunakan teknik *dimensionality reduction* atau reduksi dimensi yang bertujuan mengurangi jumlah dimensi data tanpa menghilangkan informasi penting. Salah satu metode yang banyak digunakan adalah UMAP (*Uniform Manifold Approximation and Projection*), yang bekerja dengan mempelajari struktur data berdimensi tinggi lalu memproyeksikannya ke ruang berdimensi lebih rendah (McInnes dkk., 2020). Dibandingkan dengan metode lain seperti t-SNE, UMAP dinilai lebih mampu mempertahankan struktur data secara menyeluruh, baik hubungan lokal antar kelompok data maupun hubungan global antar kelompok yang berbeda (Dorrity dkk., 2020).

2.6.3 Klastering

Klastering merupakan teknik analisis data yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan tingkat kemiripan tertentu. Dalam konteks pemodelan topik, klastering bertujuan menemukan pola atau struktur dalam kumpulan dokumen sehingga topik-topik utama yang dibahas dapat diidentifikasi secara sistematis.

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) adalah algoritma klastering berbasis kepadatan yang mampu mengidentifikasi klaster dengan tingkat kepadatan yang berbeda-beda (McInnes dkk., 2017). Algoritma ini mengelompokkan titik-titik data yang berdekatan secara spasial ke dalam klaster yang sama, tanpa mengharuskan bentuk dan ukuran klaster seragam. Selain itu, HDBSCAN mampu menangani *noise* dengan baik, sehingga cocok untuk data teks yang kompleks dan tidak terstruktur (Stewart & Al-Khassaweneh, 2022).

2.6.4 *Class-Based* TF-IDF

TF-IDF merupakan metode dalam pemrosesan teks yang digunakan untuk mengukur tingkat kepentingan suatu kata dalam dokumen atau korpus. Metode ini mengombinasikan dua komponen utama, yaitu term *frequency* (TF) dan *inverse document frequency* (IDF) (Geeksforgeeks, t.t.). Secara umum, bobot TF-IDF dirumuskan sebagai berikut:

$$W_{k,d} = tf_{k,d} \times log\,(\frac{N}{df_k}) \quad (2.1)$$

$tf_{k,d}$ = frekuensi kata k dalam dokumen d

df_k = jumlah dokumen yang mengandung kata k

N = jumlah seluruh dokumen dalam korpus

Komponen TF merepresentasikan seberapa sering kata k muncul dalam suatu dokumen, sedangkan IDF mengukur tingkat keunikan atau keinformatifan kata tersebut dalam keseluruhan korpus. Nilai IDF dihitung menggunakan logaritma dari perbandingan antara jumlah dokumen dan jumlah dokumen yang mengandung kata k.

Dalam pendekatan *class-based* TF-IDF (c-TF-IDF), prosedur TF-IDF diterapkan pada kelompok dokumen, di mana setiap klaster diperlakukan sebagai satu dokumen tunggal yang dibentuk dari penggabungan seluruh dokumen dalam klaster tersebut. Dengan pendekatan ini, bobot kata dihitung berdasarkan kepentingannya dalam suatu klaster, bukan pada dokumen individual. Rumus c-TF-IDF dapat dinyatakan sebagai berikut:

$$W_{k,c} = tf_{k,c} \times log\,(1 + \frac{A}{tf_k}) \quad (2.2)$$

$tf_{k,c}$ = frekuensi kata k dalam klaster c

tf_k = frekuensi kata k pada seluruh klaster

A = rata-rata jumlah kata per klaster

Pada formulasi ini, term *frequency* menggambarkan seberapa sering kata k muncul dalam klaster c, sedangkan *inverse class frequency* digunakan untuk menilai seberapa informatif kata tersebut bagi suatu klaster dibandingkan klaster lainnya. Penambahan konstanta 1 bertujuan untuk memastikan nilai logaritma tetap positif.

Dengan demikian, c-TF-IDF memungkinkan pemodelan kepentingan kata pada tingkat klaster, sehingga distribusi kata yang terbentuk lebih representatif untuk menggambarkan

topik. Selain itu, representasi topik dapat disederhanakan lebih lanjut dengan menggabungkan topik yang jarang muncul ke dalam topik yang lebih umum sesuai kebutuhan analisis.

2.7 Optimasi Pemodelan Topik

Meskipun pendekatan pemodelan topik berbasis *embedding* kontekstual seperti BERTopic menawarkan keunggulan dalam menangkap makna semantik teks pendek, tetap saja terdapat tantangan teknis. Tantangan utama berasal dari karakteristik data media sosial X. Subbab ini mengidentifikasi tantangan-tantangan utama tersebut serta menegaskan kebutuhan optimasi parameter agar topik yang dihasilkan bersifat koheren, stabil, dan representatif terhadap wacana publik.

Pemodelan topik berbasis *embedding* dalam BERTopic melibatkan dua tahapan kritis, yaitu reduksi dimensi menggunakan *Uniform Manifold Approximation and Projection* (UMAP) dan klasterisasi dokumen menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Kualitas topik yang dihasilkan sangat dipengaruhi oleh konfigurasi parameter pada kedua tahapan ini.

2.7.1 Sensitivitas Parameter UMAP

UMAP berfungsi memproyeksikan *embedding* berdimensi tinggi ke ruang berdimensi lebih rendah dengan tetap mempertahankan struktur data. Salah satu parameternya adalah *n_neighbors.* Parameter *n_neighbors* menentukan sejauh mana struktur lokal atau global data dipertahankan. Nilai *n_neighbors* yang kecil cenderung menekankan struktur lokal sehingga menghasilkan klaster yang lebih granular, sementara nilai yang besar menekankan struktur global dan menghasilkan klaster yang lebih umum. Pemilihan parameter yang tidak tepat dapat berdampak langsung pada kualitas topik yang terbentuk (McInnes dkk., 2020).

2.7.2 Sensitivitas Parameter HDBSCAN

Tahap klasterisasi dalam BERTopic dilakukan menggunakan HDBSCAN, yang dirancang untuk mengidentifikasi klaster berbasis kepadatan sekaligus memisahkan *noise*. Salah satu parameternya adalah *min_cluster_size.* Parameter *min_cluster_size* menentukan ukuran minimum klaster yang dianggap valid. Nilai *min_cluster_size* yang terlalu kecil berpotensi menghasilkan banyak klaster kecil yang kurang bermakna, sedangkan nilai yang terlalu besar dapat menggabungkan klaster yang seharusnya terpisah atau mengklasifikasikan sebagian besar dokumen sebagai *noise*. Proporsi *noise* yang terlalu tinggi dibandingkan

ekspektasi domain dapat mengindikasikan salah satunya dari miskonfigurasi parameter (McInnes dkk., 2017).

2.7.3 Interaksi Kompleks antar Parameter

Parameter UMAP dan HDBSCAN saling memengaruhi dalam pembentukan struktur klaster. Kombinasi tertentu dari *n_neighbors* dan *min_cluster_size* dapat menghasilkan struktur klaster yang sangat berbeda meskipun *dataset* yang digunakan sama. Studi sebelumnya menunjukkan bahwa perbedaan konfigurasi parameter dapat menghasilkan variasi yang signifikan pada kualitas topik (Syahrial & Afidh, 2024). Oleh karena itu, pemilihan parameter yang tidak sistematis berisiko menghasilkan topik yang tidak stabil dan sulit direproduksi.

2.7.4 Pendekatan Optimasi

Pemilihan parameter secara manual atau berbasis intuisi tidak akan efisien. Pendekatan optimasi yang sistematis diperlukan untuk mengeksplorasi kombinasi parameter secara efisien dan terukur. Studi yang dilakukan oleh Wijanto dkk. (2024) menunjukkan bahwa *hyperparameter tuning* berperan penting dalam mengidentifikasi konfigurasi model yang paling sesuai. Studi lainnya juga menunjukkan bahwa optimasi parameter melalui *hyperparameter tuning* secara eksplisit menghasilkan peningkatan yang signifikan pada koherensi dan stabilitas topik dibandingkan konfigurasi bawaan atau *trial-and-error* (Syahrial & Afidh, 2024).

Diperlukan sebuah pendekatan untuk menerapkan *hyperparameter tuning* dalam mengonfigurasi parameter. Pendekatan seperti *Grid search* maupun *random search* menjadi kurang efisien pada skala parameter yang besar (Andika & Kusrini, 2025). Disisi lain, Optuna dipandang sebagai metode yang relevan karena mampu melakukan pencarian kombinasi hyperparameter secara adaptif dan sistematis, serta menghindari keterbatasan pendekatan manual (Srinivas & Katarya, 2022). Oleh karena itu, penelitian ini menerapkan strategi optimasi dengan *hyperparameter tuning* menggunakan Optuna untuk memastikan topik yang dihasilkan merepresentasikan wacana publik secara lebih akurat dan dapat dipertanggungjawabkan secara metodologis.

2.8 Penelitian Terdahulu

Pada subbab ini akan dibahas penelitian-penelitian terdahulu yang relevan dengan pemodelan topik untuk analisis data teks sebagai landasan konseptual dan metodologis penelitian ini. Penelitian terdahulu pada kajian ini difokuskan pada studi-studi yang menerapkan pemodelan topik untuk menganalisis data teks. Penelusuran literatur dilakukan melalui Google Scholar dengan menggunakan kata kunci "*Topic Modeling*", "Pemodelan Topik Latent Dirichlet Allocation", dan "Pemodelan Topik BERTopic". Literatur yang dipilih mencakup penelitian yang secara eksplisit melaporkan metode pemodelan topik, sumber data, serta metrik evaluasi model, sehingga memungkinkan dilakukan perbandingan metodologis dan analisis kinerja.

Tabel 2.1 Bingkai Analisis Pemodelan Topik

Sumber	Judul	Metode dan Data	Hasil
(Khadijah & Cahyono, 2024)	Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA)	LDA, data ulasan dari Google Maps.	3 topik utama memiliki rata-rata nilai *perplexity* ≈ -8.8 dan *coherence score* ≈ 0.3. Jumlah topik topik sangat minim.
(Puspita dkk., 2024)	Analisis Pemodelan Topik Ulasan Aplikasi BNI, BCA, dan BRI Menggunakan Latent Dirichlet Allocation	LDA, data ulasan aplikasi mobile banking BNI, BCA, dan BRI dari Playstore.	6 topik utama dengan nilai *coherence* ≈ 0.404. Hasil topik relevan mengenai merek Somethinc dan interpretasi topik cukup baik secara kualitatif.
(Oktafiandi, 2023)	Implementasi LDA untuk Pengelompokan Topik Twitter Bertagar #Mypertamina	LDA, data media sosial X	3 Topik dengan dengan nilai coherence ≈ 0.468. Hasil topik belum menangkap isu secara mendalam.
(Aryani dkk., 2024)	Pemodelan topik pemberitaan Pemilu 2024 pada media daring	BERTopic, artikel berita daring.	5 topik utama yang merepresentasikan fokus pemberitaan media menjelang Pemilu 2024. *Coherence* ≈ 0.0902, namun topik yang dihasilkan cukup interpretatif.
(Listyawan dkk., 2017)	Pengelompokan topik aktivitas pengembangan perangkat lunak	BERTopic, data aktivitas proyek perangkat lunak.	6 topik utama dengan *coherence* 0.625 dan *diversity* 0.828, meskipun masih terdapat tumpang tindih topik.

	menggunakan BERTopic		
(Wahyuni dkk., 2025)	Pemodelan topik wacana publik terhadap kebijakan Program Makan Bergizi Gratis	BERTopic, kolom komentar YouTube.	9 topik utama yang mengungkap apresiasi dan kekhawatiran publik terhadap kebijakan Makan Bergizi Gratis. *Coherence score* sebesar 0,46 dan *topic diversity* sebesar 76%,

Berdasarkan Tabel 2.1, terlihat bahwa pendekatan *Latent Dirichlet Allocation* (LDA) masih banyak digunakan pada penelitian yang menganalisis ulasan daring dan data media sosial. Meskipun LDA mampu menghasilkan sejumlah topik utama dengan nilai *coherence* yang relatif moderat, sebagian besar penelitian tersebut menunjukkan keterbatasan dalam hal interpretabilitas topik. Nilai *coherence* yang rendah pada beberapa studi mengindikasikan bahwa topik yang dihasilkan cenderung kurang representatif secara semantik. Penelitian oleh Oktafiandi (2023) dan juga Khadijah & Cahyono (2024) menghasilkan jumlah topik yang sangat sedikit. Hal ini menunjukkan topik-topik yang tersembunyi tidak dapat dilihat.

Sebaliknya, penelitian yang menggunakan BERTopic menunjukkan kecenderungan menghasilkan topik yang lebih interpretatif pada beragam jenis data. Studi Aryani dkk. (2024) menunjukkan bahwa meskipun nilai *coherence* relatif rendah, topik yang dihasilkan tetap mampu merepresentasikan fokus wacana media secara tematik. Hasil yang lebih kuat ditunjukkan oleh Listyawan dkk. (2017) serta Wahyuni dkk. (2025), yang melaporkan nilai *coherence* dan *topic diversity* yang lebih tinggi, sehingga struktur topik yang terbentuk lebih beragam dan informatif. Temuan ini mengindikasikan bahwa pendekatan berbasis *embedding* seperti BERTopic lebih adaptif terhadap variasi konteks dan karakteristik data dibandingkan metode probabilistik klasik.

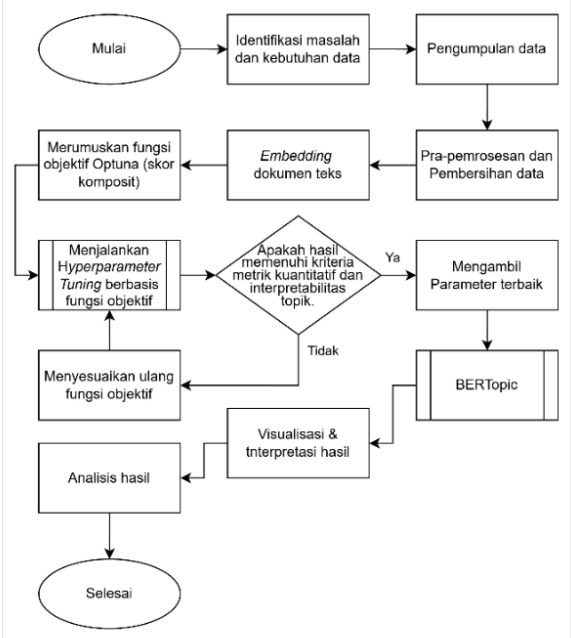
BAB III
METODOLOGI PENELITIAN

3.1 Desain Metodologis dan Reprodusibilitas Penelitian

Bab ini menjelaskan kerangka metodologis penelitian yang digunakan untuk memodelkan struktur topik mengenai Danantara di platform media sosial X. Penelitian ini menggunakan pendekatan kuantitatif berbasis analisis teks dengan kerangka *neural topic modeling* dari BERTopic. Pendekatan tersebut diintegrasikan dengan optimasi parameter reduksi dimensi UMAP dan klasterisasi HDBSCAN untuk memperoleh struktur topik yang koheren dan stabil serta interpretatif.

Fokus utama Bab ini diarahkan pada perancangan metodologis, meliputi alur teknis pemrosesan data, strategi optimasi parameter berbasis fungsi objektif, serta mekanisme evaluasi kualitas topik. Ruang lingkup pembahasan teknis mencakup pengumpulan dan pembersihan data, pembentukan representasi semantik menggunakan *embedding* berbahasa Indonesia, reduksi dimensi dan klasterisasi, serta pemodelan topik dan optimasi parameter hingga diperoleh model final.

Kerangka metodologi penelitian secara keseluruhan disajikan pada Gambar 3.1, yang menggambarkan alur pelaksanaan penelitian dari tahap pengumpulan data hingga tahap finalisasi model secara sistematis dan terukur. Diagram tersebut menjadi acuan utama dalam penyusunan struktur Bab III, sehingga setiap subbab mengikuti alur dan urutan aktivitas metodologis yang ditampilkan dalam alur penelitian.



Gambar 3.1 Diagram Alir Penelitian

Dokumentasi setiap tahapan penelitian menjadi aspek penting untuk menjamin kredibilitas dan reprodusibilitas hasil. Seluruh proses eksperimen dijalankan menggunakan perangkat lunak dan pustaka yang terdefinisi secara eksplisit, yaitu BERTopic versi 0.17.3, UMAP versi 0.5.9.post2, HDBSCAN versi 0.8.40, serta Optuna versi 4.5.0. Seluruh alur pemrosesan data, konfigurasi parameter model, dan hasil optimasi terdokumentasi dalam repositori sebagai berikut https://github.com/akhnafal-aban/Danantara-Research. Selain itu, data penelitian, ringkasan hasil *tuning* parameter, serta keluaran statistik utama dari proses pemodelan disediakan secara terpisah melalui media penyimpanan daring yang dirujuk dalam dokumentasi repositori (*README.md)* dalam repositori tersebut. Penyediaan artefak penelitian ini dimaksudkan untuk memastikan transparansi proses, memungkinkan verifikasi independen, serta mendukung replikasi penelitian pada konteks dan data yang serupa.

3.2 Pendefinisian Kebutuhan dan Analisis Awal

Penelitian ini menggunakan pendekatan kuantitatif berbasis analisis teks untuk mengidentifikasi topik lalu memetakan struktur wacana publik mengenai Danantara di platform media sosial X. Pendekatan ini dipilih karena data yang dianalisis berupa teks pendek dalam jumlah besar yang bersifat tidak terstruktur dan *noisy*. Oleh karena itu, fokus pembahasan akan diarahkan pada pendefinisian kebutuhan metodologis yang relevan agar struktur diskursus publik dapat dimodelkan secara representatif dan terukur.

Pendefinisian kebutuhan penelitian berangkat dari keterbatasan pendekatan analisis sentimen. Yang hanya mengandalkan klasifikasi polaritas emosi dan berisiko mereduksi kompleksitas wacana. Untuk memperoleh gambaran awal mengenai karakteristik wacana dan korpus yang dianalisis, penelitian ini melakukan pemodelan topik menggunakan data pengambilan awal dengan konfigurasi bawaan dan tanpa optimasi parameter. Tahap eksplorasi ini dimaksudkan sebagai sarana diagnosis awal terhadap struktur data, pola wacana, serta potensi artefak pemodelan yang muncul.

Berdasarkan evaluasi terhadap hasil eksplorasi yang dilakukan, kebutuhan penelitian kemudian dirumuskan sebagai kebutuhan untuk membangun suatu alur pemrosesan melingkup pra-pemrosesan dan pembersihan. Tahap eksplorasi awal mengindikasikan adanya karakteristik khas data media sosial, seperti aktivitas *burst*, duplikasi konten, serta potensi distorsi akibat aktivitas tidak organik. Hal ini didapatkan melalui hasil topik awal yang cenderung positif dan sama sekali tidak ada topik yang berbau kritik. Sehingga, dapat disimpulkan perlu adanya pendekatan untuk mendeteksi aktivitas *burst* tersebut.

Selain *burst,* kebutuhan dalam mengoptimasi parameter juga penting. Hasil eksplorasi selanjutnya menunjukkan topik yang dihasilkan melalui parameter bawaan sering kali menghasilkan topik yang sangat umum dan dominan. Sehingga, diperlukan pendekatan untuk mengonfigurasi parameter secara optimal menggunakan *hyperparameter tuning* dengan Optuna.

Hasil dari pemodelan topik awal tersebut juga menunjukkan bahwa terdapat topik yang menghasilkan kata kunci tidak bermakna. Kata kunci ini meliputi kata-kata yang biasanya dihapus dalam penelitian pemodelan topik yang disebut *stopwords.* Oleh karena itu, pemodelan topik menggunakan BERTopic tetap diperlukan penghapusan *stopwords*.

Untuk selanjutnya, diperlukannya juga kebutuhan untuk menentukan indikator yang menandakan keberhasilan penelitian. Keberhasilan penelitian ini ditentukan oleh kombinasi indikator kuantitatif dan kualitatif yang secara bersama-sama merefleksikan kualitas pemodelan topik. Pendekatan ini digunakan karena evaluasi pemodelan topik yang bersifat eksploratif tidak dapat sepenuhnya direduksi ke dalam metrik numerik saja tanpa mempertimbangkan aspek interpretabilitas topik (Meaney dkk., 2023).

Secara kuantitatif, indikator utama yang digunakan adalah *topic coherence* dan *topic diversity. Topic coherence* digunakan untuk mengukur keterkaitan semantik antar kata dalam satu topik, sedangkan *topic diversity* digunakan untuk mengukur tingkat perbedaan antartopik agar menghindari terbentuknya topik yang saling tumpang tindih (R. Li dkk., 2023). Kedua

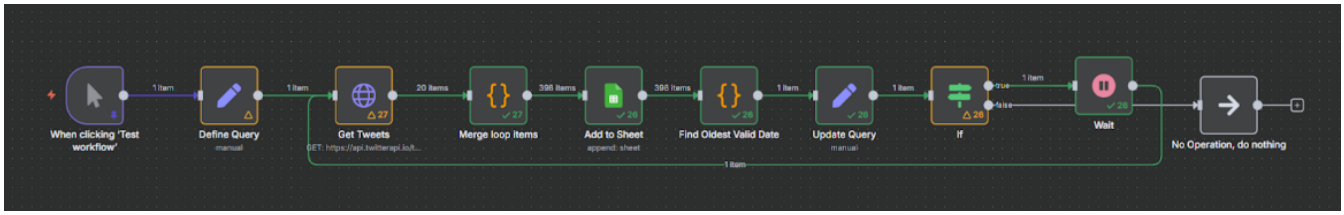
metrik ini umum digunakan dalam evaluasi pemodelan topik (Grootendorst, 2022; Riaz dkk., 2025). Sementara itu, penelitian ini tidak mengasumsikan bahwa nilai metrik kuantitatif yang tinggi secara otomatis menghasilkan topik yang bermakna. Oleh karena itu, penilaian kualitatif juga ditetapkan sebagai indikator keberhasilan penelitian.

Penilaian secara kualitatif dilakukan melalui evaluasi interpretabilitas topik, yaitu sejauh mana kata kunci dan dokumen (cuitan) dalam satu topik dapat dipahami sebagai satu isu tematik yang koheren (Meaney dkk., 2023). Keterkaitan semantik antar kata kunci utama akan ditelaah dengan relevansi beberapa dokumen representatif yang memiliki probabilitas lebih dari 0.9 terhadap topik terkait. Dengan mengombinasikan evaluasi kuantitatif dan kualitatif, topik yang dihasilkan diharapkan tidak hanya optimal secara metrik, tetapi juga relevan dan bermakna dalam memetakan struktur wacana publik secara analitis.

3.3 Proses Pengumpulan dan Deskripsi Hasil Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa teks percakapan publik yang dipublikasikan pada platform media sosial X. Data utama yang diambil adalah konten teks cuitan dengan *metadata* numerik seperti waktu publikasi, *URL*, dan jumlah interaksi. *Metadata* tersebut dimanfaatkan untuk keperluan pembersihan data, pengendalian duplikasi, serta analisis yang berkaitan fenomena *burst*.

Proses pengumpulan data dilakukan secara otomatis dengan memanfaatkan *workflow* automasi n8n yang dapat dilihat pada Gambar 3.2, yang berfungsi untuk mengatur alur pengambilan data, penjadwalan eksekusi, dan penyimpanan hasil *scraping* secara konsisten serta terautomasi. Automasi ini memungkinkan pengambilan data dilakukan secara berulang dan terkontrol. Dalam setiap eksekusi *workflow*, n8n mengelola permintaan data, menangani respons API, serta menyimpan hasil pengambilan ke dalam Google Sheet.



Gambar 3.2 *Workflow* n8n

Sumber data percakapan publik diperoleh melalui layanan twitterapi.io, yang menyediakan akses ke data cuitan publik berdasarkan kueri pencarian tertentu. Layanan ini digunakan sebagai perantara untuk mengekstraksi data dari platform X karena menawarkan biaya yang lebih murah dibandingkan antarmuka pemrograman aplikasi resmi. Penggunaan

twitterapi.io difokuskan pada pengambilan konten teks dan *metadata* dasar yang sudah dijelaskan sebelumnya.

Pengumpulan data dilakukan dalam dua tahap pengambilan yang terpisah. Tahap pertama dilakukan pada rentang Januari hingga Maret yang ditujukan sebagai penyediaan data awal untuk kebutuhan analisis dan eksplorasi pemodelan topik. Keragaman data pada tahap pertama ini dinilai belum memadai untuk mendukung pemodelan topik yang komprehensif. Oleh karena itu, tahap kedua dilakukan pada bulan November untuk memperluas cakupan data dan memperkaya variasi wacana yang dianalisis. Data dari kedua tahap tersebut kemudian digabungkan menjadi satu korpus yang lebih besar.

Proses pengambilan data dibatasi menggunakan kueri pencarian spesifik, yaitu `danantara lang:id since:2025-03-23 -from:grok`. Pembatasan bahasa `lang:id` digunakan untuk memastikan data hanya dalam lingkup berbahasa Indonesia, sementara parameter waktu `since:2025-03-23` berfungsi mengontrol rentang temporal pada salah satu tahap pengambilan data. Pengecualian terhadap akun tertentu `-from:grok` diterapkan untuk mengurangi konten otomatis atau nonnaratif yang berpotensi mengganggu pembentukan struktur topik. Pembatasan kueri ini ditujukan untuk meningkatkan relevansi data terhadap penelitian dan bukan untuk menyaring atau mengarahkan jenis opini tertentu.

Aspek etika penelitian dijaga dengan membatasi penggunaan identitas pengguna. Seluruh data diperlakukan secara anonim. Seluruh data yang dihasilkan disimpan dalam format CSV.

3.4 Pemrosesan dan Pemodelan Topik Berbasis BERTopic

Subbab ini menjelaskan *pipeline* teknis utama penelitian yang digunakan untuk membentuk struktur topik awal dari korpus percakapan publik mengenai Danantara. *Pipeline* ini bersifat deterministik dan berurutan, di mana setiap tahapan menghasilkan keluaran yang menjadi masukan bagi tahapan berikutnya. Tujuan utama pada tahap ini adalah membangun struktur pemodelan topik dasar yang konsisten secara metodologis, sebelum dilakukan optimasi parameter pada tahap selanjutnya. Seluruh proses pada subbab ini menggunakan konfigurasi parameter bawaan dan belum melibatkan mekanisme optimasi berbasis fungsi objektif.

3.4.1 Pra-pemrosesan Teks

Proses pra-pemrosesan meliputi normalisasi dasar teks, seperti penghapusan karakter nonalfanumerik yang tidak relevan, serta pembersihan spasi berlebih. Tautan, *mention*, dan simbol tertentu yang tidak berkontribusi pada makna semantik dihapus untuk mengurangi *noise*. Selain itu, penghapusan *stopword* diterapkan secara selektif dengan mengombinasikan daftar *stopword* bahasa Indonesia dari pustaka *nltk* yang daftarnya dapat dilihat di raniaprillyap (2020) dan daftar *stopword* khusus yang disusun berdasarkan observasi awal terhadap korpus yang penulis telah lampirkan pada Bab Lampiran . Penyusunan *stopword* khusus ini didasarkan pada hasil topik awal pemodelan pada Gambar 3.3, yang menunjukkan bahwa sejumlah kata kunci *stopwords* atau token umum membentuk klaster topik tersendiri yang tidak merepresentasikan isu tematik layaknya topik yang lain.

19	17	3242	17_bumn_danantara_udah_logo
20	18	2545	18_investasi_danantara_pasar_investor
21	19	562	19_iki_opo_ae_ora
22	20	2426	20_jabatan_penasehat_pejabat_dewan
3	1	6170	1_ekonomi_indonesia_danantara_dukung
4	2	3671	2_prabowo_presiden_badan_investasi
5	3	9763	3_danantara_udah_gue_kalo
6	4	1877	4_bpi_ekonomi_energi_danantara

Gambar 3.3 Hasil Pemodelan Topik Awal

Beberapa topik secara dominan tersusun oleh token percakapan informal, seperti `udah`, `gue`, `kalo`, serta `iki`, `opo`, `ae`, dan `ora`. Dominasi token-token tersebut menyebabkan topik yang terbentuk lebih merepresentasikan pola bertutur pengguna daripada struktur wacana kebijakan yang dianalisis, sehingga memperkuat kebutuhan penerapan *stopword* khusus dalam konteks penelitian ini.

3.4.2 Pembersihan Data dan Reduksi *Noise*

Setelah tahap pra-pemrosesan teks, dilakukan pembersihan data. Tahap ini difokuskan pada dua aspek utama, yaitu penghapusan duplikasi data dan pengurangan *noise* berbasis anomali *burst* temporal. Pembersihan lanjutan ini didasarkan pada temuan awal tentang topik umum yang dominan dan minim kritik.

a. Penghapusan Duplikasi Data

Penghapusan duplikasi dilakukan untuk memastikan bahwa setiap dokumen dalam korpus merepresentasikan satu percakapan yang unik. Duplikasi diidentifikasi dengan menjadikan kolom *URL* sebagai kunci utama. Pendekatan ini efektif untuk menghilangkan duplikasi yang muncul akibat perbedaan pengambilan data lintas periode, maupun kegagalan mekanisme automasi pada tahap awal penyusunan *workflow.*

Selain berbasis *URL*, pemeriksaan kemiripan juga dilakukan pada kolom teks cuitan (*cleanedTweets*) yang sudah melewati tahap pra-pemrosesan, untuk mendeteksi konten spam atau cuitan identik yang memiliki *URL* berbeda. Kombinasi identifikasi berbasis *URL* dan teks diterapkan untuk menjaga kualitas korpus dan mencegah penggandaan dokumen yang dapat mendistorsi distribusi topik pada tahap pemodelan.

b. Pengurangan *Noise* Berbasis Anomali *Burst* Temporal

Selain duplikasi, korpus media sosial rentan terhadap *noise* yang muncul akibat pola produksi konten serupa yang berulang dalam interval waktu yang sangat rapat (Singh & Kumari, 2021). Dalam konteks penelitian ini, *burst* temporal didefinisikan sebagai pola kemunculan suatu term yang terkonsentrasi secara signifikan dalam rentang waktu yang sangat singkat, sehingga membentuk lonjakan frekuensi yang tidak sebanding dengan distribusi kemunculan term tersebut sepanjang periode pengamatan. Fenomena *burst* temporal dapat mendistorsi hasil pemodelan topik karena lonjakan kemunculan term dalam waktu singkat berpotensi menutupi pola percakapan yang lain. Akibatnya, topik-topik yang mencerminkan dinamika diskursus publik secara alami menjadi kurang terlihat karena didominasi oleh term yang diproduksi secara masif dalam interval waktu terbatas. Oleh karena itu, penelitian ini menerapkan tahap pengurangan *noise* berbasis deteksi anomali *burst* temporal sebagai langkah sebelum pemodelan, dengan tujuan membatasi pengaruh percakapan yang terlalu mendominasi dan tidak natural.

Deteksi *burst* temporal dalam penelitian ini dilakukan dengan menitikberatkan pada konsentrasi kemunculan term dalam interval waktu yang sempit yaitu. Tahap ini mencakup analisis frekuensi dan distribusi temporal terhadap token, meliputi tagar dan *bigram* (frasa), yang diekstraksi setelah proses tokenisasi dan normalisasi tanpa menghapus tagar. Setiap term dihitung frekuensi globalnya dan diurutkan untuk mengidentifikasi token yang muncul secara tidak proporsional dalam korpus. Peringkat frekuensi ini digunakan sebagai dasar untuk memfokuskan analisis temporal pada term-term yang berpotensi mendominasi wacana secara struktural.

Selanjutnya, dilakukan analisis distribusi temporal untuk setiap term dengan mengurutkan seluruh cuitan yang memuat term tersebut berdasarkan stempel waktu publikasi. Selisih waktu antarposting dihitung untuk mengidentifikasi interval kemunculan yang tidak lazim. Ambang batas *burst* ditetapkan secara dinamis menggunakan kuantil distribusi global selisih waktu, sehingga definisi *burst* mencerminkan karakteristik temporal spesifik dari korpus yang dianalisis. Berdasarkan ambang tersebut, *burst* didefinisikan sebagai rangkaian kemunculan term dengan selisih waktu antarpublikasi yang berada di bawah nilai kuantil tersebut.

Setiap *burst* yang terbentuk kemudian dievaluasi menggunakan tiga metrik keabnormalan temporal. Metrik pertama adalah *Burst Size Ratio* (BSR) pada persamaan (3.1), yang dihitung sebagai perbandingan antara ukuran *cluster burst* terbesar dengan frekuensi total kemunculan term. Nilai BSR yang tinggi menunjukkan bahwa sebagian besar kemunculan term terkonsentrasi pada satu *burst* tunggal, suatu pola yang jarang ditemukan pada percakapan organik.

$$BSR = |C_{Max}| \div \Sigma_i f_i \quad (3.1)$$

Secara matematis, BSR dihitung sebagai perbandingan antara ukuran klaster *burst* terbesar (C_{Max}) dengan total frekuensi kemunculan term tersebut pada seluruh periode pengamatan ($\Sigma_i f_i$). Nilai BSR yang tinggi menunjukkan bahwa sebagian besar kemunculan term terkonsentrasi pada satu *burst* tunggal, yang mengindikasikan pola kemunculan yang tidak merata secara temporal dan relatif jarang ditemukan pada percakapan yang bersifat organik dan stabil. Sebaliknya, nilai BSR yang rendah mencerminkan distribusi kemunculan term yang lebih tersebar dari waktu ke waktu, sehingga lebih konsisten dengan dinamika percakapan publik yang alami.

Metrik kedua adalah *Repeated Burst Count* (RBC) yang ditunjukkan pada persamaan (3.2), yang digunakan untuk mengukur jumlah *burst* bermakna yang muncul secara berulang dengan ukuran minimal tertentu ($size_{Min}$). Dalam penelitian ini, nilai $size_{Min}$ ditetapkan sebesar 20 untuk menyaring *burst* dengan signifikansi yang memadai dan mengurangi *noise* dari fluktuasi temporal berskala kecil. Secara matematis, RBC dihitung sebagai jumlah klaster *burst* yang memiliki ukuran sama dengan atau lebih besar dari ambang batas tersebut. Nilai RBC yang tinggi mengindikasikan adanya pola kemunculan term yang berulang dan terstruktur dalam rentang waktu pengamatan, yang berpotensi mencerminkan dinamika penyebaran informasi yang terkoordinasi atau dipicu oleh peristiwa tertentu.

$$RBC = \Sigma_j I(|C_j| \geq size_{Min}) \quad (3.2)$$

Dalam perumusan metrik kedua tersebut, simbol Σ_j menunjukkan proses penjumlahan atas seluruh indeks klaster *burst* j yang ada dalam data, sedangkan fungsi indikator I bernilai satu apabila ukuran klaster burst ke-j ($|C_j|$) memenuhi atau melampaui ambang batas ukuran minimum ($size_{Min}$), dan bernilai nol apabila kondisi tersebut tidak terpenuhi. Dengan demikian, setiap klaster burst yang memenuhi kriteria ukuran minimum akan berkontribusi satu satuan terhadap nilai RBC, sementara klaster yang lebih kecil tidak diperhitungkan. Nilai RBC yang tinggi merefleksikan keberadaan banyak *burst* bermakna yang muncul secara berulang, yang mengindikasikan pola kemunculan term yang terstruktur secara temporal dan berpotensi tidak bersifat acak.

Metrik ketiga adalah *burst duration mean* (BDM), yang menghitung rata-rata durasi *burst* dalam satuan menit dan digunakan sebagai parameter interpretatif untuk memahami kepadatan temporal, namun tidak dijadikan indikator penentu keabnormalan karena sensitivitasnya terhadap variasi ekstrem.

Sebuah *burst* dikategorikan sebagai sumber *noise* apabila memenuhi sedikitnya dua dari tiga kriteria operasional berikut:

a. Frekuensi token global ≥ 2000
b. Nilai BSR ≥ 0,30
c. Nilai RBC ≥ 5.

Penetapan nilai batas pada ketiga kriteria dilakukan berdasarkan analisis awal terhadap karakteristik data, bukan berdasarkan ambang baku tertentu. Ambang frekuensi token global ≥ 2000 ditetapkan karena ditemukan sejumlah term dengan lonjakan kemunculan yang sangat tidak proporsional, bersifat sangat dominan positif, dan jauh melampaui term negatif maupun netral. Pola ini mengindikasikan produksi konten yang tidak natural dan berpotensi mendistorsi struktur wacana.

Nilai BSR ≥ 0,30 dan RBC ≥ 5 ditentukan melalui pengamatan terhadap term-term yang secara semantik bersifat umum dan berulang dalam berbagai pembahasan. Term dengan nilai tersebut cenderung berkontribusi besar terhadap lonjakan temporal tanpa membawa makna tematik yang spesifik, sehingga lebih berfungsi sebagai penguat volume percakapan daripada representasi isu substantif.

Sebuah *burst* dikategorikan sebagai *noise* apabila memenuhi minimal dua dari tiga kriteria untuk menghindari penghapusan berbasis satu indikator tunggal. Seluruh cuitan dalam *burst* terkait kemudian dihapus agar struktur topik yang dihasilkan tidak terdistorsi oleh konten yang diproduksi secara tidak alami. Evaluasi terpisah atas masing-masing indikator tidak dilakukan karena berada di luar ruang lingkup penelitian, dan pembersihan data difokuskan pada peningkatan stabilitas dan kualitas pemodelan topik.

3.4.3 Pembentukan *Embedding* dengan IndoSBERT

Setelah melalui tahap pra-pemrosesan dan pembersihan, setiap dokumen teks ditransformasikan ke dalam bentuk representasi numerik menggunakan IndoSBERT sebagai model *embedding* berbahasa Indonesia (Diana & Khodra, 2023). Proses ini mengubah korpus teks menjadi matriks *embedding* berdimensi tetap melalui kode sebagai berikut:

```
embeddings = SentenceTransformer("denaya/indoSBERT-
large").encode(texts, batch_size=128, show_progress_bar=True,
convert_to_numpy=True)
```

Gambar 3.4 Kode Pembentukan Embedding

Proses pembentukan *embedding* dilakukan dengan mendefinisikan model IndoSBERT-large melalui fungsi `SentenceTransformer` dari *library* `sentence-transformers`. Model ini digunakan sebagai *encoder* untuk mengubah setiap dokumen teks menjadi representasi numerik. Selanjutnya, proses *encoding* dilakukan menggunakan metode `encode()` dengan parameter `batch_size` bernilai 128, yang berfungsi untuk membagi korpus teks ke dalam kelompok-kelompok kecil agar proses komputasi lebih efisien. Parameter `show_progress_bar=True` digunakan untuk menampilkan progres proses *encoding*. Hasil *encoding* kemudian dikonversi ke dalam format *array NumPy* melalui parameter `convert_to_numpy=True`, agar *embedding* yang dihasilkan dapat disimpan untuk digunakan secara langsung dan menghemat proses komputasi. Dengan demikian, *embedding* ini menjadi fondasi utama dalam seluruh tahapan pemodelan topik.

3.4.4 Reduksi Dimensi Menggunakan UMAP

Embedding dokumen yang dihasilkan memiliki dimensi tinggi dan bersifat padat, sehingga kurang optimal untuk proses klasterisasi langsung. Oleh karena itu, tahap berikutnya adalah reduksi dimensi menggunakan *Uniform Manifold Approximation and Projection* (UMAP). UMAP digunakan untuk memproyeksikan *embedding* berdimensi tinggi ke ruang

berdimensi lebih rendah dengan tetap mempertahankan struktur semantik data. Reduksi dimensi ini berfungsi untuk mempermudah proses klasterisasi dengan mengurangi kompleksitas ruang vektor, sekaligus mempertahankan kedekatan semantik antar dokumen.

3.4.5 Klasterisasi Dokumen Menggunakan HDBSCAN

Hasil reduksi dimensi dari UMAP kemudian digunakan sebagai masukan untuk proses klasterisasi dokumen menggunakan *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Algoritma ini dipilih selain karena bawaan dari BERTopic sendiri tapi juga karena kemampuannya mengidentifikasi klaster berbasis kepadatan sekaligus memisahkan dokumen yang tidak termasuk dalam klaster mana pun sebagai *noise*. HDBSCAN membentuk kelompok dokumen yang memiliki kedekatan semantik tinggi dalam ruang hasil reduksi dimensi. Dokumen yang tidak memenuhi kriteria kepadatan minimum diklasifikasikan sebagai *noise* dan tidak dipaksakan masuk ke dalam klaster tertentu. Proses klasterisasi ini menghasilkan struktur klaster yang merepresentasikan kelompok percakapan dengan kemiripan makna, namun belum diinterpretasikan sebagai topik secara eksplisit.

3.4.6 Pembentukan Topik Menggunakan BERTopic melalui c-TF-IDF

Tahap selanjutnya dalam *pipeline* ini diarahkan pada proses pembentukan topik sebagai tahap akhir pemodelan dasar. Pada tahap ini, klaster dokumen yang dihasilkan oleh HDBSCAN diinterpretasikan sebagai topik melalui ekstraksi kata-kata representatif menggunakan pendekatan *class-based* TF-IDF (c-TF-IDF). Pendekatan ini memungkinkan identifikasi kata kunci yang secara relatif paling membedakan satu klaster dari klaster lainnya berdasarkan distribusi kata dalam masing-masing klaster.

Hasil dari tahap ini berupa sekumpulan topik yang masing-masing direpresentasikan oleh daftar kata kunci dan himpunan dokumen terkait. Tahap pemodelan awal menunjukkan bahwa proporsi dokumen yang diklasifikasikan sebagai *outlier* relatif tinggi. Fenomena ini sejalan dengan karakteristik HDBSCAN dalam BERTopic yang secara eksplisit memisahkan dokumen dengan kepadatan rendah sebagai *noise*. Meskipun mekanisme ini berperan penting dalam menjaga kemurnian klaster, proporsi *outlier* yang terlalu besar berpotensi mengurangi keterwakilan struktur wacana dalam analisis lanjutan.

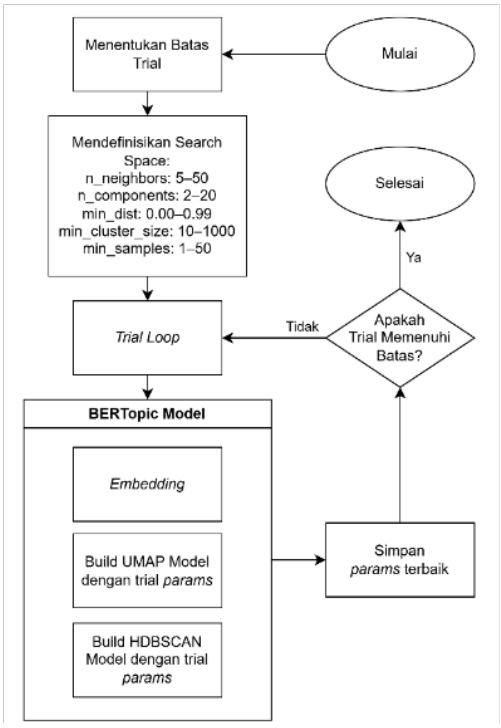
Oleh karena itu, diterapkan mekanisme *outlier reduction* yang disediakan secara bawaan dalam pustaka BERTopic untuk mereduksi jumlah *outlier* dan mendistribusikan sebagian

dokumen ke topik terdekat berdasarkan probabilitasnya. Prosedur ini tidak mengubah struktur klaster utama, melainkan berfungsi untuk meningkatkan cakupan representasi dan distribusi topik untuk kepentingan analisis.

Dengan demikian, Subbab ini menyajikan alur pemodelan topik dasar, mulai dari pembentukan *embedding* hingga pembentukan topik. Tahap ini menjadi landasan metodologis sebelum dilakukan peningkatan kualitas struktur topik melalui optimasi parameter UMAP dan HDBSCAN yang dibahas secara khusus pada subbab berikutnya.

3.5 Optimasi Parameter UMAP dan HDBSCAN melalui *Hyperparameter Tuning*

Optimasi dilakukan melalui pendekatan *hyperparameter tuning* yang terstruktur dan iteratif dengan memanfaatkan Optuna sebagai kerangka kerja optimasi. Tahapan ini secara iteratif berperan langsung dalam menentukan kualitas struktur topik yang dihasilkan. Diagram alir proses optimasi parameter dirujuk pada Gambar 3.5.



Gambar 3.5 Diagram Alir *Hyperparameter Tuning* dengan Optuna

3.5.1 Ruang Pencarian Parameter

Optimasi dilakukan pada parameter-parameter utama UMAP dan HDBSCAN yang secara langsung memengaruhi pembentukan kualitas topik. Untuk UMAP, parameter yang dioptimasi meliputi `n_neighbors`, `n_components`, dan `min_dist`. Parameter

`n_neighbors` mengontrol keseimbangan antara pelestarian struktur lokal dan global data, `n_components` menentukan dimensi ruang hasil reduksi, sedangkan `min_dist` mengatur kepadatan distribusi titik dalam ruang proyeksi (McInnes dkk., 2020). Untuk HDBSCAN, parameter yang dioptimasi meliputi `min_cluster_size` dan `min_samples`. Parameter `min_cluster_size` menentukan ukuran minimum klaster yang dianggap valid, sementara `min_samples` mengontrol sensitivitas algoritma terhadap *noise* dan kepadatan klaster (McInnes dkk., 2017).

Ruang pencarian parameter ditetapkan dalam rentang nilai sebagai berikut:

a. n_neighbors: 5–50
b. n_components: 2–20
c. min_dist: 0.0–0.99
d. min_cluster_size: 10–1000
e. min_samples: 1–50

Rentang nilai tersebut ditetapkan berdasarkan eksplorasi awal dan dokumentasi resmi *library umap-learn* dan *hdbscan* (McInnes, t.t.-a, t.t.-b). Pertama, untuk `n_neighbors` (5–50) dipilih untuk menjaga keseimbangan kritikal antara struktur lokal dan global. Berdasarkan dokumentasi UMAP, nilai yang terlalu rendah ($n < 5$) cenderung menghasilkan *embedding* yang terfragmentasi akibat fokus berlebihan pada *noise* lokal, sedangkan nilai yang terlalu tinggi ($n > 50$) akan mengaburkan variasi detail antar-topik karena algoritma terlalu memprioritaskan struktur global data. Rentang 5 hingga 50 adalah zona stabilitas yang memungkinkan algoritma menangkap nuansa topik spesifik tanpa kehilangan konteks hubungan antar-dokumen.

Kedua, untuk parameter `n_components` (2–20), penetapan ini melampaui sekadar kebutuhan visualisasi (2D/3D). McInnes dkk. (2020) menekankan bahwa UMAP tidak memiliki batasan komputasi pada dimensi *embedding* dan sering kali performa klasterisasi meningkat pada dimensi yang sedikit lebih tinggi dari 3. Batas atas 20 ditetapkan untuk memberikan ruang bagi representasi fitur laten yang lebih kaya tanpa memicu *curse of dimensionality* yang dapat mengurangi efektivitas HDBSCAN dalam mengukur densitas.

Ketiga, `min_dist` (0.0–0.99) mengatur tingkat pemadatan titik dalam *embedding*. Nilai rendah menghasilkan struktur yang sangat rapat dan kondusif untuk *clustering*, sedangkan nilai tinggi menekankan struktur global dengan sebaran lebih longgar. Rentang penuh ini sesuai yang ada di dalam dokumentasi dan memungkinkan eksplorasi spektrum *embedding* secara

menyeluruh, yang krusial karena *min_dist* sering berdampak signifikan pada separasi *cluster* dan kualitas topik.

Keempat, untuk parameter `min_cluster_size` (10–1000), rentang ini ditetapkan berdasarkan definisi operasional mengenai signifikansi topik dalam konteks analisis media sosial. Dalam dokumentasi HDBSCAN, parameter ini disebut sebagai penentu utama pembentukan klaster (McInnes, t.t.-b). Pada batas bawah 10, meskipun *default* HDBSCAN mengizinkan nilai 5, batas bawah 10 dipilih untuk menyaring mikro-topik. Dalam data Twitter yang *noisy*, klaster dengan kurang dari 10 dokumen sering kali hanya berupa repetisi spam atau anomali lokal yang tidak merepresentasikan narasi publik yang substansial. Pada batas atas 1000, batas ini didasarkan pada eksplorasi awal yang menunjukkan terdapat term yang mengarah ke isu tertentu yang jumlahnya ribuan. Jika *dataset* mengandung puluhan ribu cuitan, nilai 1000 memungkinkan algoritma mengidentifikasi *macro-topics* yang mendominasi percakapan tanpa terpecah menjadi variasi-variasi kecil yang berlebihan. Rentang yang lebar (10-1000) memberikan keleluasaan bagi Optuna untuk menentukan apakah data lebih baik direpresentasikan oleh banyak topik spesifik atau sedikit topik besar berdasarkan densitas aktual data.

Kelima, untuk parameter `min_samples` (1–50), rentang ini difokuskan pada kontrol konservatisme algoritma terhadap *noise*. Parameter ini menentukan seberapa "skeptis" HDBSCAN dalam memasukkan sebuah titik ke dalam klaster; semakin tinggi nilainya, semakin banyak titik yang akan dibuang sebagai *outlier* (McInnes, t.t.-b). Pada batas bawah 1, mengizinkan skenario *clustering* yang sangat "rakus", di mana hampir semua data dipaksa masuk ke dalam topik terdekat. Ini relevan jika tujuan analisis adalah memaksimalkan *topic coverage* meskipun berisiko menurunkan koherensi. Mengingat karakteristik data Twitter yang penuh *noise*, batas atas 50 memberikan ruang bagi algoritma untuk membuang dokumen yang benar-benar tidak memiliki kedekatan semantik yang kuat dengan inti topik.

Perlu ditekankan bahwa penetapan ruang pencarian parameter ini merupakan batasan heuristik yang dirancang berdasarkan karakteristik data tekstual media sosial. Rentang ini tidak dimaksudkan untuk mencakup seluruh kemungkinan matematis dari algoritma UMAP dan HDBSCAN, melainkan dibatasi pada zona eksplorasi yang relevan untuk ekstraksi topik. Dengan demikian, ruang pencarian ini merepresentasikan kompromi yang terukur antara efisiensi komputasi pencarian parameter dan kedalaman analisis struktur data.

3.5.2 Kerangka Hyperparameter Tuning Berbasis Optuna

Proses optimasi parameter dilakukan menggunakan Optuna sebagai kerangka kerja *hyperparameter tuning*. Optuna digunakan untuk mengelola eksplorasi ruang parameter melalui serangkaian *trial*, di mana setiap *trial* merepresentasikan satu kombinasi parameter UMAP dan HDBSCAN yang dihasilkan melalui evaluasi bertahap oleh Optuna (Srinivas & Katarya, 2022).

Dalam setiap *trial*, Optuna menjalankan *pipeline* pemodelan topik yang terdiri atas penggunaan *embedding* dokumen, reduksi dimensi menggunakan UMAP, klasterisasi menggunakan HDBSCAN, dan ekstraksi topik. Seluruh konfigurasi parameter dan hasil evaluasi disimpan melalui berkas CSV untuk mendukung analisis dan seleksi konfigurasi yang terbaik berdasarkan hasil skor yang dihasilkan oleh perumusan fungsi objektif.

3.5.3 Perumusan Fungsi Objektif

Perumusan fungsi objektif dalam optimasi hyperparameter BERTopic dilakukan secara iteratif berdasarkan evaluasi pada setiap percobaan dengan menggunakan indikator kuantitatif dan kualitatif yang telah ditetapkan. Fungsi objektif pertama kali dirumuskan secara sederhana menggunakan metrik *coherence* untuk mengamati struktur topik awal (Syahrial & Afidh, 2024). Hasil dari percobaan dievaluasi dan digunakan sebagai dasar untuk menyesuaikan formulasi fungsi objektif pada tahap berikutnya. Penyesuaian fungsi objektif dilakukan dengan ketentuan sebagai berikut:

a. Penambahan komponen evaluasi, dilakukan ketika fungsi objektif sebelumnya belum mampu mengendalikan aspek seperti jumlah topik, distribusi ukuran klaster, atau pemisahan antartopik.
b. Penerapan komponen pembatas atau penalti, digunakan untuk mengurangi kecenderungan model yang menghasilkan konfigurasi ekstrem.
c. Penyesuaian bobot antar komponen, dilakukan untuk menyeimbangkan nilai masing-masing metrik dalam fungsi objektif. Hal ini mencegah dominasi satu metrik yang dapat menurunkan nilai aspek lainnya.

Melalui siklus evaluasi dan penyesuaian yang berulang, fungsi objektif secara bertahap diarahkan untuk menghasilkan konfigurasi model tinggi secara numerik namun juga dapat diinterpretasikan secara analitis.

Setiap percobaan dicatat dalam dua berkas CSV yang berbeda. Berkas pertama memuat nilai metrik evaluasi yang digunakan, nilai total fungsi objektif, konfigurasi parameter yang

digunakan, dan jumlah topik yang dihasilkan pada setiap *trials* yang dibuat oleh Optuna. Berkas kedua memuat informasi topik yang dihasilkan melalui hasil konfigurasi dengan nilai total fungsi objektif tertinggi, meliputi kata kunci utama dan representasi dokumen. Pencatatan ini memungkinkan analisis dan evaluasi terhadap topik yang dihasilkan seperti apa dan verifikasi konsistensi fungsi objektif sebelum konfigurasi tersebut digunakan sebagai acuan pada percobaan berikutnya.

3.6 Finalisasi Model

Model final ditetapkan berdasarkan hasil optimasi *hyperparameter* dengan capaian fungsi objektif paling optimal. Paling optimal dimaksudkan dengan mempertimbangkan keseimbangan antara hasil metrik dan penilaian kualitatif. Pertimbangan tersebut digunakan untuk memastikan bahwa konfigurasi yang dipilih tidak hanya unggul secara teknis, tetapi juga stabil dan layak digunakan sebagai dasar analisis wacana publik.

Setelah konfigurasi optimal ditetapkan, parameter terbaik digunakan untuk menjalankan pemodelan topik secara final. Model yang dihasilkan kemudian disimpan sebagai model terbaik dan dijadikan acuan pada seluruh tahap analisis selanjutnya.

3.7 Validasi Interpretasi Topik Menggunakan *Human Annotator*

Setelah model final menghasilkan keluaran topik, peneliti terlebih dahulu melakukan interpretasi topik secara mandiri. Interpretasi ini mencakup penentuan makna topik berdasarkan kata kunci dan dokumen representatif yang dihasilkan model, tanpa campur tangan pihak lain. Tujuan tahap ini adalah untuk membangun pemahaman awal mengenai struktur wacana dan karakteristik setiap topik dari perspektif peneliti.

Setelah interpretasi mandiri selesai, langkah berikutnya adalah memastikan bahwa interpretasi topik yang dihasilkan dapat dipahami dan representatif secara manusiawi. Validasi dilakukan dengan melibatkan *human annotator* sebagai *validator* interpretasi. *Annotator* menilai apakah kata kunci dan dokumen representatif setiap topik sesuai dengan makna dan konteks yang diharapkan, sehingga validitas interpretasi dapat diverifikasi. Kemudian, memvalidasi juga apakah interpretasi topik tersebut akan mengarah pada analisis isu-isu tematik yang bermakna sesuai dengan tujuan penelitian. Validasi ini dilakukan melalui instrumen Google Form sebagai berikut https://s.id/c8r8i, yang memuat skala penilaian dan masukkan untuk komentar kualitatif, sehingga interpretasi dapat diverifikasi dan diperkuat secara empiris.

Skala penilaian dalam instrumen validasi menggunakan *linear scale* dengan rentang nilai 1–5, di mana nilai 1–2 merepresentasikan penilaian negatif, nilai 3 bersifat netral, dan nilai 4–5 merepresentasikan penilaian positif. Hanya interpretasi dengan nilai 1–2 yang wajib mengisi alasan ketidaksesuaian interpretasi terhadap kata kunci dan dokumen representatif. Instrumen Google Form ini terdiri atas tiga bagian utama. Bagian A memuat identitas *human annotator* serta pernyataan persetujuan partisipasi. Bagian B digunakan untuk memvalidasi interpretasi masing-masing topik berdasarkan kesesuaian antara kata kunci, dokumen representatif, dan interpretasi yang disusun peneliti. Selanjutnya, Bagian C difokuskan pada penilaian apakah keseluruhan topik dan interpretasinya telah memadai untuk mengarah pada pencapaian tujuan penelitian.

Hasil penilaian *annotator* kemudian digunakan untuk menyesuaikan atau memperkuat interpretasi awal. Topik-topik yang memperoleh skor relevansi tinggi dan koherensi konsisten (≥ 3) dipertahankan, sedangkan topik yang menunjukkan ketidaksesuaian atau ambiguitas (≤ 3) dianalisis ulang dengan mempertimbangkan komentar *annotator*. Pendekatan ini secara keseluruhan bertujuan untuk meminimalkan bias dan lingkup asumsi interpretasi yang hanya bersumber dari peneliti, dengan melibatkan pihak independen yang memiliki latar belakang keilmuan sesuai dengan konteks penelitian, khususnya bidang politik dan bahasa, sehingga

interpretasi topik tidak hanya konsisten dengan bukti empiris dalam data, tetapi juga dapat dipahami dan dipertanggungjawabkan secara kontekstual.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Ruang Lingkup Pembahasan

Ruang lingkup pembahasan dibatasi pada penyajian hasil dari setiap tahap pengolahan data dan pemodelan. Tidak terdapat penambahan metode ataupun asumsi analitis di luar yang telah ditetapkan. Seluruh hasil yang dipaparkan disusun bertahap mengikuti alur transformasi data. Alur tersebut dimulai dari karakteristik data mentah, dilanjutkan dengan proses pembersihan dan reduksi *noise*, hingga pemodelan topik menggunakan *hyperparameter tuning* dengan Optuna. Dengan demikian, setiap temuan yang muncul pada tahap lanjutan selalu dapat ditelusuri kembali ke kondisi data pada tahap sebelumnya.

4.2 Karakteristik *Dataset* Hasil Pengumpulan

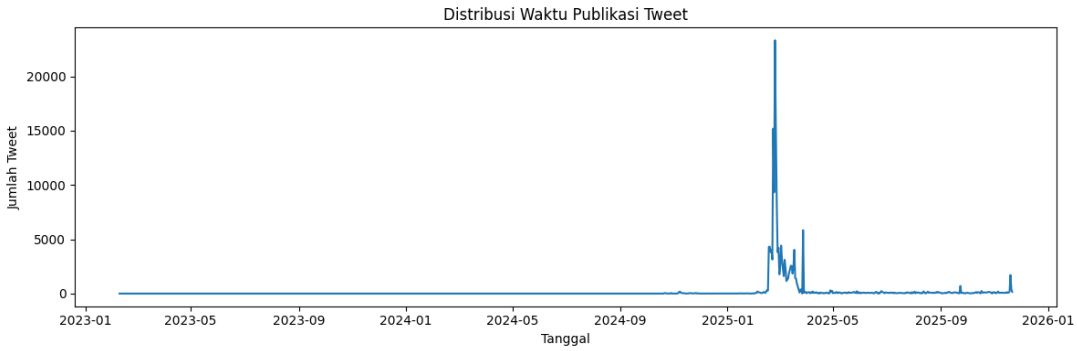
Dataset yang dianalisis merupakan hasil pengumpulan cuitan publik terkait Danantara di platform X dengan total 177.793 entri. Setiap entri data yang digunakan memuat informasi dasar berupa teks cuitan, URL, serta waktu publikasi. Pada tahap ini, data masih berada dalam kondisi mentah.

	url	createdAt	tweets
count	177793	177790	177793
unique	134034	NaN	132494
top	https://x.com/grok/status/1990965248380342689	NaN	The smartest AI just got an upgrade.\n\nTry Gr...
freq	892	NaN	917
mean	NaN	2025-03-20 00:55:57.582991616	NaN
min	NaN	2023-02-09 18:00:00	NaN
25%	NaN	2025-02-23 09:43:00	NaN
50%	NaN	2025-02-26 06:40:00	NaN
75%	NaN	2025-03-14 10:18:00	NaN
max	NaN	2025-11-21 06:20:00	NaN

Gambar 4.1 Statistik Deskriptif *Dataset* Mentah

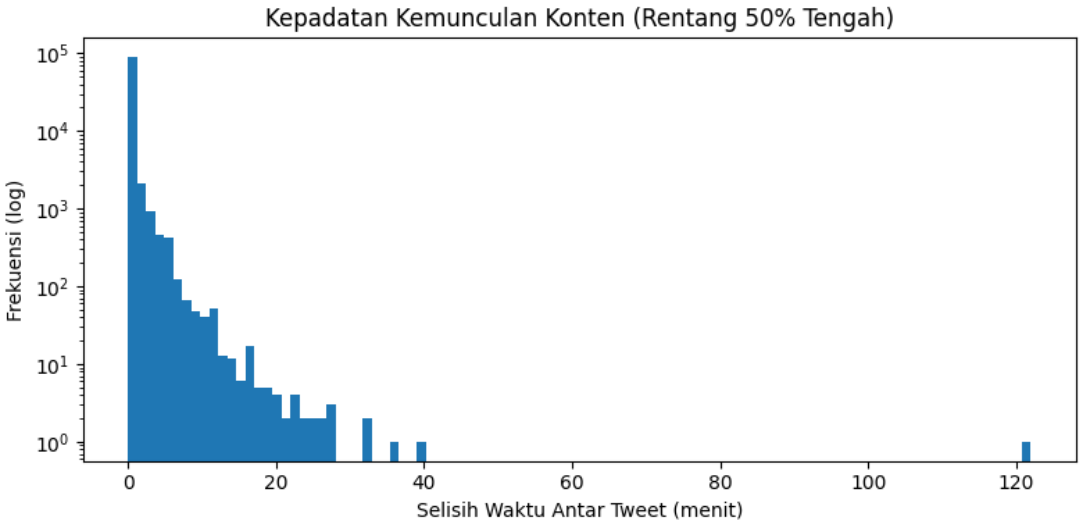
Statistik deskriptif awal yang disajikan pada Gambar 4.1 menunjukkan bahwa dari 177.793 URL, terdapat 134.034 URL unik, sementara kolom teks cuitan mencatat 132.494 entri unik. Perbedaan antara jumlah total dan jumlah unik tersebut menunjukkan adanya kemunculan konten yang berulang dalam *dataset*. Informasi waktu publikasi cuitan awal tercatat pada 9 Februari 2023 dan cuitan paling akhir pada 21 November 2025. Nilai rata-rata

waktu publikasi berada pada 20 Maret 2025, dengan 50% dari total data berada dalam rentang waktu 23 Februari 2025 hingga 14 Maret 2025. Ringkasan statistik ini memberikan konteks kuantitatif awal untuk menelaah distribusi waktu publikasi yang bisa dilihat pada Gambar 4.2.



Gambar 4.2 Distribusi Waktu Publikasi Cuitan

Distribusi waktu publikasi yang ditampilkan pada Gambar 4.2 memperlihatkan sebaran waktu yang sangat tidak merata. Visualisasi tersebut menunjukkan variasi volume cuitan dari waktu ke waktu, yang di mana sebagian besar periode awal memiliki jumlah publikasi yang relatif rendah, sementara pada rentang waktu antara bulan Januari hingga Mei terjadi peningkatan jumlah cuitan dalam interval yang lebih rapat. Pola sebaran ini selanjutnya menjadi landasan untuk mengamati kepadatan kemunculan konten berdasarkan selisih waktu antar cuitan yang bisa dilihat pada Gambar 4.3.



Gambar 4.3 Kepadatan kemunculan Konten

Kepadatan kemunculan konten dianalisis melalui selisih waktu antar cuitan berurutan dan divisualisasikan pada Gambar 4.3. Visualisasi histogram tersebut menggambarkan

distribusi selisih waktu antar publikasi dalam satuan menit, dengan analisis dibatasi pada rentang interkuartil temporal 23 Februari hingga 14 Maret 2025, yaitu periode dengan konsentrasi data tertinggi sebagaimana ditunjukkan pada distribusi waktu publikasi sebelumnya. Sumbu horizontal merepresentasikan jeda waktu antar cuitan, sementara sumbu vertikal (Frekuensi (log)) menunjukkan frekuensi kejadian selisih waktu tersebut dalam skala logaritmik. Hal ini dikarenakan jika ditampilkan tanpa skala logaritmik, akan menghasilkan distribusi frekuensi yang berat sebelah.

Distribusi pada Gambar 4.3 memperlihatkan konsentrasi frekuensi yang tinggi pada selisih waktu yang sangat pendek. Pola ini menunjukkan bahwa sebagian besar cuitan dipublikasikan dalam interval waktu yang rapat, sementara sebagian kecil lainnya memiliki jeda waktu yang jauh lebih panjang. Penyajian ini menggambarkan tingkat intensitas publikasi yang tinggi pada fase paling aktif dari rentang waktu tersebut. Hal ini menyiratkan secara tidak langsung adanya aktivitas yang tidak natural.

Dalam konteks platform X, dinamika percakapan publik seringkali tercermin melalui penggunaan frasa berulang, serta tagar yang bisa dilihat sebagai daftar *Trending Topic*. Oleh karena itu, setelah mengamati kepadatan waktu publikasi, eksplorasi distribusi frekuensi token yang meliputi frasa dan tagar menjadi relevan sebagai langkah analisis lanjutan untuk memahami bentuk keseragaman konten yang muncul dalam periode dengan intensitas tinggi.

Tabel 4.1 Distribusi Frekuensi Token

No	Term	Frekuensi	No	Term	Frekuensi
1	#DanantaraPorosInvestasi	18580	12	#tepatsolutifindonesia	307
2	#DanantaraBangunNegeri	17708	13	#presidenprabowo	294
3	danantara terpercaya	13595	14	#berita	282
4	danantaraterpercaya	13595	15	#bumn	219
5	#GerakCepatEkonomiHebat	2270	16	#IndonesiaGelap2025	214
6	#Danantara	2174	17	#Prabowo	209
7	#IndonesiaGelap	1540	18	#IndonesiaMaju	192
8	#danantara	997	19	#IndonesiaEmas	183
9	#BuruhBersamaPrabowo	621	20	#AdiliJokowi	172
10	#investasi	343	21	#KaburAjaDulu	168
11	#DanantaraUntukIndonesia	320	22	#prabowo	157

Distribusi frekuensi token pada Tabel 4.1 menunjukkan bahwa kemunculan token tertentu dalam korpus mentah bersifat sangat besar. Sejumlah kecil token seperti `#DanantaraPorosInvestasi`, `#DanantaraBangunNegeri`, dan `danantara terpercaya` memiliki frekuensi kemunculan yang jauh lebih tinggi dibandingkan token lainnya, sementara sebagian besar token muncul dalam jumlah yang relatif rendah. Pola ini memperlihatkan adanya dominasi token dalam korpus.

Seluruh temuan pada bagian ini disajikan sebagai indikasi awal berbasis statistik yang terdeskripsi, yang berfungsi sebagai landasan empiris untuk memasuki tahapan pembersihan lanjutan. Tahapan tersebut akan menangani duplikasi konten dan anomali *burst* temporal.

4.3 Pembersihan Data dan Deteksi Anomali Temporal

Berdasarkan karakteristik *dataset* mentah yang telah dipaparkan sebelumnya, khususnya terkait kemunculan konten berulang, kepadatan waktu publikasi yang tinggi, serta dominasi sejumlah token tertentu, dilakukan tahap pembersihan data untuk mengurangi distorsi topik. Pembersihan ini mencakup dua aspek utama, yaitu penghapusan duplikasi konten dan reduksi anomali *burst* temporal, yang bertujuan menghasilkan *dataset* akhir yang lebih representatif untuk tahap analisis lanjutan.

Sebelum dilakukan pembersihan data, dilakukan penambahan kolom *cleanedTweets* yang merepresentasikan teks cuitan setelah melalui pra-pemrosesan dasar, seperti normalisasi karakter, penghapusan elemen nonteks, dan penyeragaman format. Kode pra-pemrosesan bisa dilihat di Gambar 4.4. Kolom tersebut digunakan sebagai representasi teks yang lebih konsisten untuk keperluan deteksi. Dengan penambahan kolom tersebut, proses pembersihan tidak hanya mempertimbangkan kesamaan URL, tetapi juga keseragaman isi teks setelah dinormalisasi.

```
import re
def clean_tweet_text(text):
    t = str(text)
    t = t.replace("&amp;", "&").replace("&gt;",
">").replace("&lt;", "<")
    t = re.sub(r'https?://\S+|www\.\S+|t\.co/\S+', '', t)
    t = re.sub(r'@\w+', '', t)
    t = re.sub(r'\1', t)
    t = re.sub(r'([!?.,]){2,}', r'\1', t)
    t = re.sub(r'[\r\n]+', ' ', t)
    t = re.sub(r'\s+', ' ', t).strip()
```

```
    t = re.sub(r'\s+([,?.!])', r'\1', t)
    return t
merged_df['cleanedTweets'] =
merged_df['tweets'].apply(clean_tweet_text)
```

Gambar 4.4 Isi Kode pra-pemrosesan

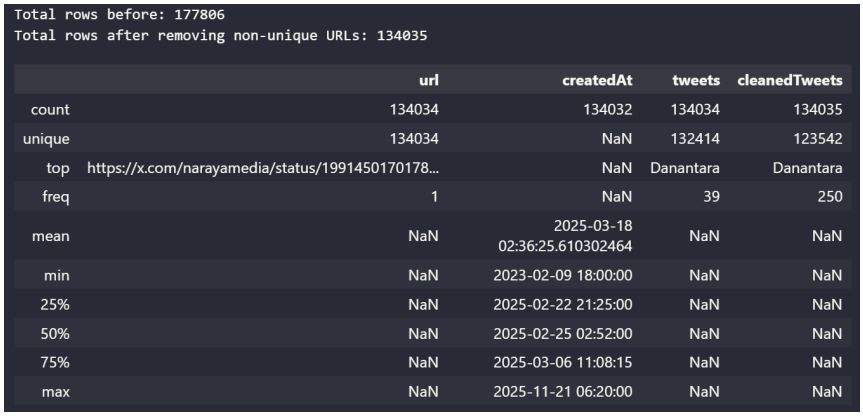
Setelah tahapan pra-pemrosesan diterapkan pada korpus utama, diperoleh perbandingan cuitan sebelum dan sesudah proses tersebut, sebagaimana ditunjukkan pada Tabel 4.2 berikut.

Tabel 4.2 Perbandingan Sebelum dan Sesudah Pra-pemrosesan Data

Sebelum	Sesudah
Danantara menjadi langkah besar dalam menciptakan sistem manajemen aset yang lebih transparan, efisien, dan berdampak bagi pembangunan nasional. #DanantaraUntukIndonesia #EkonomiKuat #IndonesiaMaju https://t.co/zYMrwPE2vT"	Danantara langkah menciptakan sistem manajemen aset transparan, efisien, berdampak pembangunan nasional. #DanantaraUntukIndonesia #EkonomiKuat #IndonesiaMaju

4.3.1 Penghapusan Duplikasi dan Spam Data

Penghapusan duplikasi dilakukan dengan mengacu pada kesamaan URL sebagai identitas utama entri. Hasil proses ini ditunjukkan pada Gambar 4.5, yang memperlihatkan bahwa jumlah URL unik telah sepenuhnya identik dengan jumlah total URL, yaitu 134.034 entri. Kondisi ini menegaskan bahwa seluruh duplikasi berbasis URL telah berhasil dihilangkan dari *dataset*.


Total rows before: 177806
Total rows after removing non-unique URLs: 134035

	url	createdAt	tweets	cleanedTweets
count	134034	134032	134034	134035
unique	134034	NaN	132414	123542
top	https://x.com/narayamedia/status/1991450170178...	NaN	Danantara	Danantara
freq	1	NaN	39	250
mean	NaN	2025-03-18 02:36:25.610302464	NaN	NaN
min	NaN	2023-02-09 18:00:00	NaN	NaN
25%	NaN	2025-02-22 21:25:00	NaN	NaN
50%	NaN	2025-02-25 02:52:00	NaN	NaN
75%	NaN	2025-03-06 11:08:15	NaN	NaN
max	NaN	2025-11-21 06:20:00	NaN	NaN

Gambar 4.5 Statistik *Dataset* Setelah Deduplikasi

Namun demikian, statistik pada Gambar 4.5 juga menunjukkan bahwa jumlah teks unik pada kolom *tweets* dan *cleanedTweets* masih lebih rendah dibandingkan jumlah URL unik.

Kolom *cleanedTweets* mencatat 123.542 entri unik, sementara total baris tetap berada pada 134.035 entri. Selisih ini menunjukkan bahwa masih terdapat sejumlah cuitan dengan isi teks yang identik atau sangat serupa setelah dinormalisasi, meskipun berasal dari URL dan entri yang berbeda.

Perbedaan tersebut mengindikasikan keberadaan konten berulang, di mana teks yang sama dipublikasikan berulang kali oleh akun yang sama atau akun yang terkoordinasi ataupun melalui mekanisme unggahan ulang dengan cara salin-tempel. Kondisi konten berulang tersebut didefinisikan sebagi konten spam. Dengan demikian, meskipun deduplikasi berbasis URL telah menyelesaikan penghapusan duplikasi, *dataset* masih mengandung pola keseragaman konten yang tidak dapat diidentifikasi melalui URL semata yaitu spam.

Untuk menangani keseragaman konten tersebut, dilakukan pembersihan lanjutan terhadap spam dengan memanfaatkan kolom *cleanedTweets*. Statistik *dataset* setelah pembersihan spam disajikan pada Gambar 4.6. Gambar tersebut menunjukkan bahwa jumlah total entri pada *dataset* menjadi 123.541 baris, dengan jumlah URL unik, teks cuitan unik, dan teks hasil normalisasi yang sepenuhnya identik, masing-masing berjumlah 123.541 entri.

	url	createdAt	tweets	cleanedTweets
count	123541	123541	123541	123541
unique	123541	NaN	123541	123541
top	https://x.com/narayamedia/status/1991450170178...	NaN	Wakil Kepala BGN Nanik S. Deyang menegaskan an...	Wakil Kepala BGN Nanik S. Deyang menegaskan an...
freq	1	NaN	1	1
mean	NaN	2025-03-18 16:58:09.390404608	NaN	NaN
min	NaN	2023-04-05 09:53:00	NaN	NaN
25%	NaN	2025-02-22 20:10:00	NaN	NaN
50%	NaN	2025-02-25 01:15:00	NaN	NaN
75%	NaN	2025-03-05 10:38:00	NaN	NaN
max	NaN	2025-11-21 06:20:00	NaN	NaN

Gambar 4.6 Statistik *Dataset* Setelah Pembersihan

Pada tahap ini, *dataset* tidak hanya bebas dari duplikasi struktural berbasis URL, tetapi juga dari *spam*. Seluruh entri yang tersisa merepresentasikan cuitan dengan URL dan konten teks yang unik. *Dataset* hasil pembersihan ini selanjutnya digunakan sebagai dasar untuk tahap reduksi anomali *burst* temporal dan analisis lanjutan.

4.3.2 Deteksi dan Reduksi *Burst* Temporal

Setelah *dataset* terbebas dari duplikasi struktural dan keseragaman konten berbasis isi, analisis dilanjutkan pada dimensi temporal untuk mengidentifikasi lonjakan kemunculan term

yang bersifat tidak proporsional dalam interval waktu tertentu. Fokus tahap ini adalah mendeteksi pola *burst*, yaitu kondisi ketika suatu term muncul secara intensif dalam periode waktu yang relatif singkat, sehingga berpotensi mendistorsi distribusi topik dalam korpus .

Berdasarkan Tabel 4.1, dominasi frekuensi global tersebut menunjukkan bahwa sebagian percakapan terkonsentrasi pada term tertentu. Namun, frekuensi tinggi tersebut belum cukup untuk menyimpulkan adanya anomali, karena frekuensi tersebut masih dapat tersebar merata sepanjang waktu. Oleh karena itu, diperlukan analisis lanjutan untuk menelusuri bagaimana kemunculan term-term dominan tersebut terdistribusi secara temporal.

Distribusi temporal kemunculan term dianalisis dengan membentuk klaster *burst* berbasis interval waktu satu menit. Ringkasan klaster *burst* yang terbentuk disajikan pada Tabel 4.3 Tabel ini memperlihatkan bahwa sejumlah term dengan frekuensi global tinggi juga membentuk klaster *burst* berukuran sangat besar dalam rentang waktu yang singkat. Sebagai contoh, `#danantaroporosinvestasi` membentuk klaster dengan ukuran 8.883 kemunculan hanya dalam beberapa jam, demikian pula `#danantarabangunnegeri` dan variasi frasa yang menyertainya.

Tabel 4.3 Ringkasan *Burst Cluster*

Term	start_time	end_time	size
#danantaraporosinvestasi	2025-02-22 15:26:00	2025-02-22 19:04:00	8883
#danantarabangunnegeri	2025-02-24 15:01:00	2025-02-24 19:05:00	5994
danantara terpercaya	2025-02-24 15:01:00	2025-02-24 19:05:00	5794
danantarabangunnegeri danantara	2025-02-24 15:01:00	2025-02-24 19:05:00	5791
#danantarabangunnegeri	2025-02-24 19:30:00	2025-02-24 21:26:00	3778
danantara terpercaya	2025-02-24 19:30:00	2025-02-24 21:26:00	3775
danantarabangunnegeri danantara	2025-02-24 19:30:00	2025-02-24 21:26:00	3774
...	...	...	...
#berita	2025-02-28 16:43:00	2025-02-28 16:49:00	15
#berita	2025-02-24 12:29:00	2025-02-24 12:35:00	10
#investasi	2025-02-24 12:29:00	2025-02-24 12:35:00	10
#presidenprabowo	2025-02-24 12:29:00	2025-02-24 12:35:00	10
#tepatsolutifindonesia	2025-02-24 12:29:00	2025-02-24 12:35:00	10
...	...	...	...

Sebaliknya, tabel yang sama juga menunjukkan keberadaan term dengan frekuensi global lebih rendah, seperti `#berita` atau `#investasi`, yang membentuk *burst* berukuran kecil dan berdurasi singkat. Perbedaan ini mengindikasikan bahwa tidak semua lonjakan temporal memiliki skala dan intensitas yang sama, sehingga diperlukan ukuran kuantitatif untuk membedakan *burst* yang bersifat ekstrem dari fluktuasi temporal yang lebih moderat.

Untuk mengkuantifikasi intensitas *burst* yang teridentifikasi, dilakukan penghitungan metrik temporal yang mencakup *burst size ratio* (BSR), *repeated burst count* (RBC), dan *burst duration mean* (BDM). Ringkasan nilai metrik untuk term-term dengan frekuensi tinggi disajikan pada Tabel 4.4. Tabel ini menunjukkan bahwa term-term tertentu memiliki kombinasi nilai frekuensi tinggi, BSR besar, serta RBC yang tinggi, yang mencerminkan lonjakan kemunculan signifikan dan pola *burst* yang terjadi berulang kali.

Tabel 4.4 Metrik Temporal Term Abnormal

Term	**Frekuensi**	**BSR**	**RBC**	**BDM**
#danantaraporosinvestasi	18602	0.4775292979	22	22.32786885
#danantarabangunnegeri	17769	0.3373290562	20	19.36486485
danantara terpercaya	13594	0.4262174489	7	16.79629628
ekonomi indonesia	5444	0.07365907421	34	6.7443609
...	...	...	...	....
masa depan	5064	0.1139415482	22	6.05128205
...	...	...	...	...
yang lebih	4386	0.06338349293	24	4.932835817
pertumbuhan ekonomi	3635	0.03768913343	21	4.726495717
bisa jadi	1437	0.03897007655	5	4.05
global transformasi	809	0.3980222497	7	14.3
...	...	...	...	...

Sebagai contoh, `#danantaroporosinvestasi` dan `#danantarabangunnegeri` tidak hanya memiliki frekuensi global yang tinggi, tetapi juga nilai BSR yang besar serta RBC di atas ambang yang telah ditetapkan, menunjukkan bahwa kemunculan term tersebut terkonsentrasi pada beberapa periode *burst* yang berulang. Sebaliknya, term seperti `masa depan` atau yang lebih memiliki frekuensi relatif tinggi namun BSR yang lebih rendah, menunjukkan bahwa kemunculannya lebih tersebar meskipun tetap membentuk *burst* berulang dalam jumlah tertentu.

Berdasarkan distribusi nilai metrik tersebut, ditetapkan kriteria abnormalitas yang menggabungkan ambang frekuensi, intensitas *burst*, dan pengulangan *burst*. Penerapan kriteria ini menghasilkan daftar term yang diklasifikasikan sebagai anomali temporal, sebagaimana dirangkum pada Tabel 4.5. Tabel ini menunjukkan bahwa term yang dihapus tidak hanya berupa tagar utama dengan frekuensi sangat tinggi, tetapi juga frasa-frasa turunan dan kombinasi term yang muncul berulang dalam pola *burst* serupa.

Tabel 4.5 Ringkasan Term yang Dihapus

Term	**Kriteria**	**Frekuensi**	**Klaster Dihapus**
#danantaraporosinvestasi	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	18097	61
#danantarabangunnegeri	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	17433	74
bpi danantara	['freq>=2000', 'RBC>=5']	1012	120
#gerakcepatekonomihebat	['freq>=2000', 'BSR>=0.30']	974	16
ekonomi indonesia	['freq>=2000', 'RBC>=5']	833	133
pertumbuhan ekonomi	['freq>=2000', 'RBC>=5']	279	117
dengan danantara	['freq>=2000', 'RBC>=5']	234	132
yang lebih	['freq>=2000', 'RBC>=5']	168	134
masa depan	['freq>=2000', 'RBC>=5']	134	117
aset negara	['freq>=2000', 'RBC>=5']	124	72
danantara bisa	['freq>=2000', 'RBC>=5']	91	45
dari danantara	['freq>=2000', 'RBC>=5']	80	76
ekonomi nasional	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	37	42
danantara jadi	['freq>=2000', 'RBC>=5']	30	77
ekonomi yang	['freq>=2000', 'RBC>=5']	15	71
danantara yang	['freq>=2000', 'RBC>=5']	14	43
bersama danantara	['freq>=2000', 'RBC>=5']	5	97
danantara terpercaya	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	3	54
danantara danantaraporosinvestasi	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	0	74
danantara danantarabangunnegeri	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	0	79

#buruhbersamaprabowo	['BSR>=0.30', 'RBC>=5']	0	14
danantarabangunnegeri buruhbersamaprabowo	['BSR>=0.30', 'RBC>=5']	0	14
danantarabangunnegeri danantara	['freq>=2000', 'BSR>=0.30', 'RBC>=5']	0	54
danantaraporosinvestasi https	['BSR>=0.30', 'RBC>=5']	0	26
ekonomi gerakcepatekonomihebat	['BSR>=0.30', 'RBC>=5']	0	16
global transformasi	['BSR>=0.30', 'RBC>=5']	0	10

Kolom jumlah klaster yang dihapus pada tabel tersebut memperlihatkan bahwa beberapa term membentuk puluhan hingga ratusan klaster *burst*, yang mengindikasikan pola lonjakan temporal pada term-term tersebut. Sebaliknya, term dengan jumlah klaster yang lebih kecil menunjukkan *burst* yang lebih terbatas meskipun tetap memenuhi kriteria abnormalitas lainnya.

Penerapan reduksi anomali temporal berdasarkan daftar term tersebut menghasilkan penghapusan 39.563 cuitan dari *dataset* pasca pembersihan spam. Setelah proses ini, korpus menyisakan 83.978 cuitan sebagaimana yang ditunjukkan di Gambar 4.7.

```
Total tweet awal: 123541
Tweet abnormal (dibuang): 39563
Tweet bersih: 83978

Persentase tweet dibuang: 32.02%
```

	url	createdAt	tweets	cleanedTweets	row_index
count	83978	83978	83978	83978	83978.000000
unique	83978	43928	83978	83978	NaN
top	https://x.com/narayamedia/status/1991450170178...	2025-02-26 04:22:00	Wakil Kepala BGN Nanik S. Deyang menegaskan an...	Wakil Kepala BGN Nanik S. Deyang menegaskan an...	NaN
freq	1	39	1	1	NaN
mean	NaN	NaN	NaN	NaN	57733.111779
std	NaN	NaN	NaN	NaN	41177.689865
min	NaN	NaN	NaN	NaN	0.000000
25%	NaN	NaN	NaN	NaN	20999.250000
50%	NaN	NaN	NaN	NaN	42574.500000
75%	NaN	NaN	NaN	NaN	101665.750000
max	NaN	NaN	NaN	NaN	123540.000000

Gambar 4.7 Statistik Pasca Reduksi Noise

Langkah reduksi ini dilakukan untuk mengurangi dominasi topik yang bersumber dari lonjakan temporal ekstrem, sehingga distribusi konten dalam korpus menjadi lebih seimbang. Sebagai penekanan, bagian ini tidak dilakukan evaluasi terhadap dampak reduksi terhadap hasil pemodelan topik, karena fokus pembahasan dibatasi pada proses penyaringan dan transformasi data sebelum tahap representasi semantik dan pemodelan.

4.3.3 *Dataset* Pasca Pembersihan Temporal

Setelah penerapan reduksi anomali temporal, korpus yang dihasilkan menyisakan 83.978 cuitan. *Dataset* ini merupakan hasil dari pembersihan duplikasi struktural, penghapusan spam berbasis isi, serta reduksi lonjakan temporal ekstrem yang berpotensi mendistorsi distribusi topik. Gambar 4.7 menunjukkan bahwa seluruh URL pada *dataset* bersifat unik dan setiap entri merepresentasikan satu dokumen cuitan yang berbeda. Namun, berdasarkan eksplorasi awal mengenai topik yang terdapat kata kunci *stopwords,* maka akan dilakukan penghapusan *stopwords* terlebih dahulu.

4.3.4 Penyempurnaan Teks melalui Penghapusan Stopwords

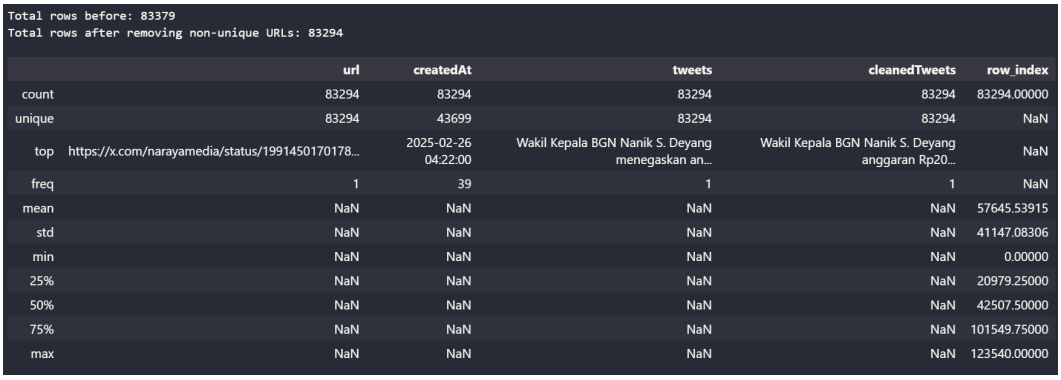
Dalam korpus saat ini, beberapa topik terbentuk hampir sepenuhnya dari kata-kata umum yang tidak membawa muatan tematik yang spesifik. Berdasarkan temuan tersebut, dilakukan penghapusan *stopwords* terhadap kolom *cleanedTweets* dengan menggunakan daftar *stopwords* bahasa Indonesia serta daftar tambahan yang disusun berdasarkan hasil eksplorasi awal topik. Dampak penghapusan *stopwords* terhadap ukuran *dataset* ditunjukkan pada Gambar 4.8.

	url	createdAt	tweets	cleanedTweets	row_index
count	83379	83379	83379	83379	83379.000000
unique	83379	43734	83379	83294	NaN
top	https://x.com/narayamedia/status/1991450170178...	2025-02-26 04:22:00	Wakil Kepala BGN Nanik S. Deyang menegaskan an...	danantara	NaN
freq	1	39	1	17	NaN
mean	NaN	NaN	NaN	NaN	57648.623694
std	NaN	NaN	NaN	NaN	41149.891178
min	NaN	NaN	NaN	NaN	0.000000
25%	NaN	NaN	NaN	NaN	20973.500000
50%	NaN	NaN	NaN	NaN	42509.000000
75%	NaN	NaN	NaN	NaN	101553.500000
max	NaN	NaN	NaN	NaN	123540.000000

Gambar 4.8 Statistik *Dataset* Setelah Penghapusan *Stopwords*

Setelah tahap ini, jumlah *dataset* berkurang dari 83.978 menjadi 83.379 cuitan. Pengurangan ini disebabkan oleh adanya sejumlah entri yang setelah penghapusan *stopwords* tidak lagi memiliki token tersisa, seperti cuitan yang hanya berisi satu kata umum atau nama entitas tunggal yang termasuk dalam daftar *stopwords*. Entri-entri tersebut dihapus karena tidak lagi merepresentasikan dokumen tekstual yang bermakna untuk analisis semantik.

Untuk memastikan bahwa setiap entri merepresentasikan satu dokumen teks yang unik, dilakukan deduplikasi ulang berbasis kolom *cleanedTweets*. Hasil deduplikasi ulang tersebut ditampilkan pada Gambar 4.9, yang menunjukkan bahwa jumlah akhir *dataset* menjadi 83.294 cuitan, dengan kesetaraan antara jumlah total baris dan jumlah teks unik.



Total rows before: 83379
Total rows after removing non-unique URLs: 83294

	url	createdAt	tweets	cleanedTweets	row_index
count	83294	83294	83294	83294	83294.00000
unique	83294	43699	83294	83294	NaN
top	https://x.com/narayamedia/status/1991450170178...	2025-02-26 04:22:00	Wakil Kepala BGN Nanik S. Deyang menegaskan an...	Wakil Kepala BGN Nanik S. Deyang anggaran Rp20...	NaN
freq	1	39	1	1	NaN
mean	NaN	NaN	NaN	NaN	57645.53915
std	NaN	NaN	NaN	NaN	41147.08306
min	NaN	NaN	NaN	NaN	0.00000
25%	NaN	NaN	NaN	NaN	20979.25000
50%	NaN	NaN	NaN	NaN	42507.50000
75%	NaN	NaN	NaN	NaN	101549.75000
max	NaN	NaN	NaN	NaN	123540.00000

Gambar 4.9 Statistik *Dataset* Akhir Pasca Deduplikasi Ulang

Dataset dengan 83.294 cuitan ini diposisikan sebagai *dataset* akhir yang digunakan pada tahap *embedding* dan pemodelan topik. Seluruh tahapan pembersihan sebelumnya memastikan bahwa korpus ini bebas dari duplikasi struktural, spam, dan lonjakan temporal abnormal.

4.4 Representasi Semantik Dokumen

Setelah melalui seluruh tahapan pembersihan dan penyempurnaan teks, korpus akhir yang terdiri dari 83.294 cuitan selanjutnya direpresentasikan ke dalam bentuk numerik melalui proses *embedding* dokumen. Representasi ini bertujuan memetakan setiap dokumen teks ke dalam ruang vektor berdimensi tetap, sehingga memungkinkan analisis berbasis jarak dan struktur semantik pada tahap berikutnya.



Preview of embeddings result and corresponding texts:

	text	embedding
0	Wakil Kepala BGN Nanik S. Deyang anggaran Rp20...	[-0.0010849195532500744, -0.747622013092041, -...
1	Hahahaha. Danantara segitu doang?	[-0.2941948175430298, 0.24598027765750885, -0....
2	The smartest AI just got an upgrade. Try Grok ...	[-0.21719172596931458, -0.6396735906600952, -0...
3	heran presiden korup otoriter eranya gelar PAH...	[0.18580776453018188, 0.20904456079006195, -0....
4	Proyek Raksasa WTE Danantara. Latar Belakang: ...	[-0.14688740670681, -0.5512070059776306, -0.55...
5	Kondisi masyarakat susah. MBG, Danantara, whoo...	[-0.13741421699523926, 0.15323136746883392, -0...
6	hehe. memuja spi kmdn main sembur kotoran kpd ...	[0.12633687257766724, -0.015082120895385742, 0...
7	Stay comfortable and stylish in our men's flee...	[0.12725140154361725, -0.31774869561195374, -0...
8	wkwkwk. jelek disalahkan partai kebaikan skrg ...	[0.2799445390701294, -0.2188941240310669, 0.07...
9	knp jokowi menumpas korupsi itu?	[-0.20303994417190552, -0.20454388856887817, 0...

Showing 10 of 83294 embeddings. Embedding dimension: 256

Gambar 4.10 Pratinjau Hasil *Embedding* dan Teks Terkait

Hasil *embedding* dokumen ditunjukkan pada Gambar 4.10. Setiap baris pada tabel tersebut merepresentasikan satu dokumen cuitan beserta vektor *embedding* yang bersesuaian.

Contoh yang ditampilkan memperlihatkan potongan kecil dari keseluruhan korpus, di mana teks cuitan yang telah dibersihkan dipasangkan dengan vektor numerik hasil transformasi.

Vektor *embedding* yang dihasilkan memiliki dimensi 256, dengan hasil vektor untuk setiap dokumen. Dengan demikian, seluruh korpus direpresentasikan sebagai matriks *embedding* berukuran 83.294 × 256, yang memuat informasi semantik laten dari masing-masing cuitan dalam bentuk numerik kontinu. Representasi ini bersifat homogen, di mana setiap dokumen memiliki panjang vektor yang sama, terlepas dari variasi panjang teks aslinya.

Dengan tersedianya representasi *embedding* untuk seluruh dokumen, *dataset* telah berada pada format yang siap digunakan untuk proses reduksi dimensi, klasterisasi, dan pemodelan topik. Subbab ini berfungsi sebagai penghubung antara tahap pembersihan data dan tahap pemodelan, dengan menegaskan bahwa seluruh dokumen talah di vektorisasi sebelum dilakukan proses tuning dan analisis topik.

4.5 *Hyperparameter tuning* BERTopic

Setelah seluruh dokumen direpresentasikan dalam bentuk *embedding* numerik, tahap selanjutnya adalah penentuan konfigurasi pemodelan topik yang paling sesuai. Proses ini dilakukan melalui serangkaian percobaan *hyperparameter tuning* menggunakan Optuna, dengan fungsi objektif dan ketentuan yang telah ditetapkan. Setiap percobaan menghasilkan konfigurasi model yang berbeda, baik dari sisi nilai metrik evaluasi maupun struktur topik yang terbentuk. Pada bagian ini, pembahasan difokuskan pada hasil kuantitatif dan analisis interpretatif, tanpa memasuki interpretasi tematik substantif.

4.5.1 Percobaan 1 – Optimasi Berbasis *Coherence*

Percobaan pertama menggunakan nilai *coherence* sebagai satu-satunya fungsi objektif dalam proses optimasi hyperparameter. Dengan demikian, fungsi objektif pada percobaan ini dirumuskan pada persamaan (3.3) sebagai berikut:

$$Skor = coherence \quad (4.3)$$

Konfigurasi terbaik diperoleh dengan parameter `n_neighbors` = 30, `n_components` = 12, `min_dist` = 0,088, `min_cluster_size` = 962, dan `min_samples` = 9. Konfigurasi tersebut menghasilkan nilai skor *coherence* sekitar 0,478.

Namun, evaluasi struktur topik berdasarkan hasil pada Tabel 4.6 memperlihatkan ketidakseimbangan distribusi dokumen yang sangat signifikan. Satu topik utama mencakup

74.598 dokumen, sementara topik lainnya berukuran jauh lebih kecil. Ketidakseimbangan ini menunjukkan bahwa optimasi berbasis *coherence* mendorong model untuk menggabungkan sebagian besar dokumen ke dalam satu klaster besar yang memiliki kata-kata umum dengan koherensi tinggi.

Tabel 4.6 Ringkasan Hasil Tuning Berbasis *Coherence*

No	Representasi Keywords	Count
1	['danantara', 'bumn', 'indonesia', 'investasi', 'negara', 'ekonomi', 'rakyat', 'prabowo', 'presiden', 'bank']	74598
2	['pertamina', 'danantara', 'pertamax', 'indonesiagelap', 'korupsi', 'isu', 'laut', 'pagar', 'oplosan', 'pertalite']	3730
3	['danantara', 'indonesia', 'triliun', 'rp', 'indonesiagelap2025', 'prabowo', 'indonesiagelap', 'investasi', 'berkahramadan2025', '2025']	2477
4	['danantara', 'mbg', 'bumn', 'program', 'dana', 'prabowo', 'uu', 'bank', 'ikn', 'proyek']	2477

Selain dominasi ukuran topik, representasi kata kunci antar topik pada Tabel 4.6 juga memperlihatkan tumpang tindih yang kuat. Kata seperti `danantara`, `bumn`, `indonesia`, `investasi`, dan `negara` muncul berulang pada lebih dari satu topik. Kondisi ini mengindikasikan bahwa meskipun *coherence* internal topik relatif tinggi, pemisahan semantik antar topik masih lemah, sehingga topik-topik yang terbentuk cenderung merepresentasikan variasi kecil dari tema yang sama.

Berdasarkan temuan tersebut, dapat disimpulkan bahwa penggunaan *coherence* sebagai fungsi objektif tunggal belum memadai untuk menghasilkan struktur topik yang terdistribusi secara seimbang dan beragam. Model cenderung memaksimalkan keterkaitan kata secara koheren, yang berimplikasi pada rendahnya kemampuan model dalam menangkap keragaman topik. Oleh karena itu, diperlukan modifikasi fungsi objektif pada percobaan selanjutnya untuk membatasi dominasi topik tunggal dan mendorong pembentukan struktur topik yang lebih terpisah.

4.5.2 Percobaan 2 – Penambahan Penalti Jumlah Topik

Berdasarkan hasil percobaan pertama, percobaan kedua dilakukan dengan menambahkan komponen penalti berbasis jumlah topik ke dalam fungsi objektif. Walaupun percobaan pertama sudah menunjukkan kurangnya keberagaman antartopik, hal tersebut akan dilakukan pada percobaan selanjutnya. Hal ini dikarenakan kebutuhan untuk memfokuskan pada

penambahan pembatas baru efektif atau tidak. Tujuan penambahan penalti berbasis jumlah topik sebagai pembatas ini adalah mengurangi kecenderungan model membentuk topik yang yang sangat besar ataupun sangat sedikit. Kode untuk menarapkan perumusan penalti ini bisa dilihat pada Gambar 4.11.

```
min_topics = 10          # batas bawah aman
min_allowed_topics = 2   # titik minimal absolut
max_topics = 20          # batas atas aman
max_allowed_topics = 40  # titik maksimal absolut
max_penalty = 0.5        # penalti maksimum

unique_topics = set(topics)
n_topics = len(unique_topics - {-1})
trial.set_user_attr('n_topics', n_topics)

penalty = 0.0

# Penalti topik terlalu sedikit
if n_topics < min_topics:
        denom = (min_topics - min_allowed_topics)
        gap = (min_topics - n_topics)
        penalty_low = max_penalty * (gap / denom)
        penalty_low = min(penalty_low, max_penalty)
        penalty += penalty_low

# Penalti topik terlalu banyak
elif n_topics > max_topics:
        denom = (max_allowed_topics - max_topics)
        gap = (n_topics - max_topics)
        penalty_high = max_penalty * (gap / denom)
        penalty_high = min(penalty_high, max_penalty)
        penalty += penalty_high

score -= penalty
```

Gambar 4.11 Kode Perumusan Penalti Jumlah Topik

Penalti dihitung berdasarkan jumlah topik yang terbentuk dengan mengecualikan label outlier (-1), sehingga penalti hanya merefleksikan struktur topik yang benar-benar terbentuk. Secara operasional, rentang jumlah topik dibagi ke dalam dua zona. Zona aman didefinisikan pada interval 10 hingga 20 topik, di mana tidak diberikan penalti atau penalti sama dengan nol.

Apabila jumlah topik berada di bawah batas bawah tersebut, penalti dikenakan secara proporsional terhadap jarak jumlah topik aktual dari batas aman, dengan titik minimal absolut ditetapkan pada 2 topik. Sebaliknya, apabila jumlah topik melebihi batas atas, penalti dihitung berdasarkan selisih terhadap batas maksimal aman dengan titik maksimal absolut pada 40 topik. Formulasi ini menghasilkan nilai penalti meningkat secara bertahap relatif terhadap jumlah topik, namun dibatasi oleh nilai maksimum sebesar 0,5. Penalti ini bertindak sebagai mekanisme regularisasi yang mendorong model untuk menentukan konfigurasi yang menghasilkan topik dengan jumlah ideal. Fungsi objektif pada percobaan ini dirumuskan sebagai:

$$Skor = coherence - penalty \quad (4.4)$$

Konfigurasi terbaik diperoleh dengan parameter `n_neighbors` = 28, `n_components` = 16, `min_dist` = 0,886, `min_cluster_size` = 352, dan `min_samples` = 22. Menghasilkan skor objektif ≈ 0,383, yang merupakan kombinasi dari nilai *coherence* ≈ 0,508 dan nilai penalti ≈ 0,125. Meskipun skor objektif keseluruhan lebih rendah dibandingkan percobaan pertama, nilai *coherence* justru meningkat. Struktur topik hasil konfigurasi ini dirangkum pada Tabel 4.7.

Tabel 4.7 Ringkasan Hasil Tuning dengan Penalti Jumlah Topik

No	Representasi Keywords	Count
1	['danantara', 'bumn', 'bank', 'uang', 'duit', 'dana', 'negara', 'masuk', 'efisiensi', 'rakyat']	34906
2	['danantara', 'ekonomi', 'indonesia', 'investasi', 'aset', 'dukung', 'negara', 'pertumbuhan', 'berkelanjutan', 'nasional']	14768
3	['danantara', 'korupsi', 'koruptor', 'orang', 'rakyat', 'kpk', 'pengawas', 'mantan', 'pejabat', 'jokowi']	12198
4	['prabowo', 'presiden', 'danantara', 'investasi', 'badan', 'pengelola', 'bpi', 'nusantara', 'bumn', 'anagata']	7162
5	['danantara', 'indonesia', 'mbg', 'triliun', 'proyek', 'tni', 'negara', 'dana', 'program', 'pemerintah']	10366
6	['pertamina', 'danantara', 'pertamax', 'isu', 'pengalihan', 'korupsi', 'indonesiagelap', 'oplosan', 'pertalite', 'laut']	3850

Evaluasi terhadap Tabel 4.7 menunjukkan bahwa jumlah topik meningkat menjadi enam topik, dengan distribusi dokumen yang lebih tersebar dibandingkan percobaan pertama. Satu

topik masih menghasilkan topik yang mempunyai jumlah dokumen relatif tinggi. Dan juga representasi kata kunci antar topik menunjukkan bahwa keberagaman masih cukup rendah. Kata-kata seperti `danantara`, `indonesia`, `negara`, dan `investasi` tetap muncul secara konsisten di hampir seluruh topik, meskipun dalam kombinasi yang berbeda.

Temuan ini menunjukkan bahwa penalti jumlah topik berhasil menambah jumlah topik, tetapi masih belum secara efektif meningkatkan keberagaman antar topik. Berdasarkan kondisi tersebut, dapat disimpulkan bahwa pengendalian jumlah topik saja belum cukup untuk mencapai struktur topik yang optimal. Diperlukan mekanisme tambahan dalam fungsi objektif yang secara eksplisit mendorong perbedaan representasi kata antar topik, bukan hanya mengatur jumlah klaster. Oleh karena itu, pada percobaan berikutnya, fungsi objektif diperluas dengan memasukkan komponen *diversity* untuk meningkatkan keberagaman dan mengurangi redundansi antar topik.

4.5.3 Percobaan 3 – Kombinasi *Coherence* dan *Diversity*

Berdasarkan hasil percobaan kedua, diketahui bahwa pengendalian jumlah topik melalui penalti berhasil memperbaiki distribusi ukuran klaster, namun belum secara memadai meningkatkan keberagaman representasi kata antar topik. Oleh karena itu, percobaan ketiga dilakukan dengan memperluas fungsi objektif melalui penambahan metrik *diversity* untuk mendorong pemisahan leksikal antar topik. Pada percobaan ini, fungsi objektif dirumuskan pada persamaan (3.5) sebagai berikut:

$$Skor = coherence + diversity - penalty \quad (4.5)$$

Konfigurasi tersebut menghasilkan parameter terbaik dengan nilai `n_neighbors` = 12, `n_components` = 10, `min_dist` = 0,429, `min_cluster_size` = 684, dan `min_samples` = 12. Hasil optimasi menunjukkan bahwa konfigurasi terbaik pada percobaan ini memperoleh skor tertinggi sebesar 1,28, dengan nilai *coherence* ≈ 0,514 dan *diversity* ≈ 0,766. Pada konfigurasi tersebut, jumlah topik yang terbentuk adalah 12 topik, yang berada dalam rentang aman sehingga penalti tidak diterapkan.

Struktur topik hasil konfigurasi terbaik percobaan ketiga disajikan pada Tabel 4.8. Tabel tersebut menunjukkan bahwa topik-topik yang terbentuk memiliki distribusi dokumen yang lebih seimbang dibandingkan percobaan sebelumnya. Tidak terdapat satu topik yang mendominasi, dan ukuran klaster tersebar pada rentang yang lebih terkendali.

Tabel 4.8 Kombinasi *coherence, diversity, dan penalty*

No	Representasi Keywords	Count
1	['danantara', 'korupsi', 'koruptor', 'orang', 'rakyat', 'kpk', 'hukum', 'bpk', 'mantan', 'pejabat']	13842
2	['ekonomi', 'danantara', 'indonesia', 'dukung', 'semoga', 'maju', 'pertumbuhan', 'semangat', 'bangsa', 'rakyat']	8715
3	['danantara', 'bumn', 'efisiensi', 'apbn', 'duit', 'uang', 'negara', 'dana', 'rakyat', 'masuk']	16211
4	['danantara', 'bahasa', 'jelasin', 'baca', 'bahas', 'bang', 'paham', 'bayi', 'gimana', 'in']	4847
5	['pertamina', 'pertamax', 'danantara', 'isu', 'indonesiagelap', 'pengalihan', 'korupsi', 'oplosan', 'laut', 'harga']	4737
6	['danantara', 'triliun', 'rp', 'indonesia', 'indonesiagelap2025', 'indonesiagelap', 'berkahramadan2025', '000', 'mbg', 'prabowo']	3282
7	['danantara', 'mbg', 'ihsg', 'gagal', 'tni', 'uu', 'ikn', 'saham', 'bumn', 'indonesia']	4999
8	['prabowo', 'presiden', 'danantara', 'jokowi', 'sby', 'dewan', 'penasehat', 'penasihat', 'indonesia', 'tepatsolutifindonesia']	5417
9	['aset', 'bumn', 'investasi', 'berkelanjutan', 'sektor', 'danantara', 'positif', 'berdampak', 'optimalisasi', 'dampak']	4086
10	['badan', 'danantara', 'investasi', 'pengelola', 'nusantara', 'prabowo', 'presiden', '2025', 'anagata', 'bpi']	5177
11	['investasi', 'indonesia', 'danantara', 'ekonomi', 'negara', 'nasional', 'pengelolaan', 'kekayaan', 'strategis', 'aset']	5886
12	['bank', 'danantara', 'bca', 'uang', 'bumn', 'mandiri', 'pindah', 'swasta', 'duit', 'nasabah']	6045

Dari sisi representasi kata kunci, Tabel 4.8 memperlihatkan bahwa penambahan *diversity* meningkatkan keberagaman antartopik. Meskipun token `danantara` masih muncul lintas topik sebagai entitas utama, kombinasi kata kunci pendamping mulai menunjukkan diferensiasi yang lebih tegas. Sebagai contoh, topik dengan kata kunci `korupsi-koruptor-kpk` terpisah dari topik `ekonomi-pertumbuhan-dukung`, serta dari topik `bahasa-jelasin-paham` yang merepresentasikan ragam percakapan informal. Pemisahan ini tidak terlihat konsisten pada percobaan pertama dan kedua.

Hasil pada Tabel 4.9 menunjukkan bahwa penambahan metrik *diversity* secara konsisten berkontribusi besar terhadap skor akhir fungsi objektif. Konfigurasi dengan skor tertinggi 1,2805 diperoleh dari kombinasi *coherence* sebesar 0,514 dan *diversity* sebesar 0,767, dengan

jumlah topik 12, yang menunjukkan bahwa *diversity* memiliki kontribusi numerik yang lebih besar dibandingkan *coherence* dalam pembentukan skor.

Tabel 4.9 Hasil Optuna *Trials* Percobaan Ketiga

Score	Coherence	Diversity	N_topics
1.280557116	0.5138904493	0.7666666667	12
1.277155862	0.4971558625	0.78	10
1.181592771	0.5065927708	0.8	9
1.074774875	0.3760248748	0.69875	320
0.848846364	0.4952749354	0.7285714286	7
0.7671036949	0	0.7671036949	839
0.5729753035	0.5729753035	1	2
0.5729753035	0.5729753035	1	2

Pola serupa juga terlihat pada beberapa konfigurasi lain dengan skor tinggi, di mana nilai *diversity* berada pada rentang 0,76–0,80, sementara nilai *coherence* bervariasi pada rentang yang lebih sempit. Bahkan, pada beberapa konfigurasi ekstrem, skor objektif yang relatif tinggi tetap dapat dicapai meskipun nilai *coherence* sangat rendah atau bahkan nol, selama nilai *diversity* sangat tinggi. Kondisi ini terlihat pada konfigurasi dengan skor 0,7671, di mana nilai *diversity* mendominasi skor akhir meskipun *coherence* bernilai nol dan jumlah topik sangat besar.

Temuan tersebut menunjukkan bahwa pada percobaan ketiga, fungsi objektif cenderung sensitif secara berlebihan terhadap komponen *diversity*. Meskipun *diversity* efektif dalam mengurangi *overlap* representasi kata antar topik dan memperjelas batas tematik, kontribusi yang terlalu dominan berpotensi mendorong konfigurasi dengan pemisahan leksikal tinggi namun stabilitas semantik yang lemah, seperti pembentukan jumlah topik yang sangat besar atau topik dengan koherensi internal yang rendah.

Dengan demikian, meskipun kombinasi *coherence* dan *diversity* terbukti meningkatkan pemisahan antar topik dibandingkan percobaan sebelumnya, distribusi nilai pada Tabel 4.9 mengindikasikan perlunya pengendalian bobot komponen *diversity* agar keseimbangan antara pemisahan leksikal dan keterkaitan semantik tetap terjaga. Oleh karena itu, pada percobaan selanjutnya dilakukan penyesuaian fungsi objektif dengan mengurangi bobot *diversity*, sehingga kontribusinya terhadap skor akhir menjadi lebih proporsional terhadap *coherence*.

4.5.4 Percobaan 4 – Final Konfigurasi

Berdasarkan evaluasi pada percobaan ketiga, percobaan keempat dilakukan dengan mengatur ulang bobot komponen *diversity* dalam fungsi objektif. Penyesuaian ini didasarkan pada temuan bahwa nilai rata-rata *diversity* pada percobaan sebelumnya relatif tinggi, sehingga kontribusinya perlu dikendalikan agar tetap proporsional terhadap *coherence*.

Pada percobaan ini, fungsi objektif dirumuskan sebagai berikut:

$$\text{Skor} = coherence + \frac{1}{2} diversity - penalty \quad (4.6)$$

Konfigurasi tersebut menghasilkan parameter terbaik dengan nilai `n_neighbors` = 14, `n_components` = 20, `min_dist` = 0,812, `min_cluster_size` = 324, dan `min_samples` = 11. Hasil optimasi menunjukkan bahwa konfigurasi terbaik memperoleh skor ≈ 0,911, dengan *coherence* ≈ 0,524 dan *diversity* ≈ 0,773. Konfigurasi ini menghasilkan 15 topik, yang berada dalam rentang aman sehingga penalti tidak diterapkan.

Tabel 4.10 Final konfigurasi

No	Representasi Keywords	Count
1	['ekonomi', 'indonesia', 'danantara', 'dukung', 'semoga', 'pertumbuhan', 'maju', 'langkah', 'rakyat', 'nasional']	8860
2	['pertamina', 'pertamax', 'isu', 'danantara', 'pengalihan', 'oplosan', 'indonesiagelap', 'korupsi', 'harga', 'pertalite']	3562
3	['korupsi', 'koruptor', 'danantara', 'kpk', 'mantan', 'bpk', 'hukum', 'audit', 'pengawas', 'diaudit']	6920
4	['danantara', 'wkwk', 'gimana', 'apa', 'kah', 'gini', 'wkwkwk', 'pak', 'eh', 'pas']	10464
5	['aset', 'berkelanjutan', 'positif', 'bumn', 'investasi', 'negara', 'sektor', 'danantara', 'optimalisasi', 'optimalkan']	2659
6	['badan', 'pengelola', 'nusantara', 'anagata', 'investasi', 'daya', 'bpi', '2025', 'presiden', 'prabowo']	3851
7	['danantara', 'orang', 'mbg', 'dana', 'pemerintah', 'rakyat', 'asing', 'negara', 'saham', 'uu']	15427
8	['bumn', 'danantara', 'apbn', 'negara', 'uang', 'masuk', 'dividen', 'rakyat', 'duit', 'dana']	9414
9	['efisiensi', 'anggaran', 'transparansi', 'danantara', 'mbg', 'transparan', 'investasi', 'hasil', 'dana', 'efisien']	3088
10	['prabowo', 'presiden', 'jokowi', 'danantara', 'sby', 'luncurkan', 'indonesia', 'presidenprabowo', 'tepatsolutifindonesia', 'berita']	3865

11	['energi', 'kerja', 'hilirisasi', 'industri', 'lapangan', 'proyek', 'danantara', 'transisi', 'investasi', 'sektor']	2053
12	['bank', 'bca', 'mandiri', 'pindah', 'danantara', 'swasta', 'duit', 'bri', 'uang', 'tabungan']	2798
13	['triliun', 'indonesiagelap', 'danantara', 'rp', 'beli', 'ndasmu', 'rakyat', 'indonesiagelap2025', 'negara', 'dana']	4325
14	['bahasa', 'bang', 'paham', 'jelasin', 'bahas', 'baca', 'danantara', 'bayi', 'penjelasan', 'tolong']	1861
15	['investasi', 'danantara', 'indonesia', 'invest', 'investor', 'global', 'ekonomi', 'strategis', 'nasional', 'sektor']	4076

Struktur topik hasil konfigurasi final disajikan pada Tabel 4.10. Tabel tersebut menunjukkan bahwa jumlah topik meningkat dibandingkan percobaan ketiga, namun tetap stabil. Tidak terdapat topik tunggal yang mendominasi secara berlebihan, dan variasi ukuran antar topik berada pada rentang yang terkendali.

Dari sisi representasi kata kunci, Tabel 4.10 memperlihatkan keberagaman yang lebih matang. Topik-topik terkait kebijakan ekonomi, isu korupsi, energi dan industri, perbankan, serta percakapan informal muncul sebagai klaster yang terpisah dengan kombinasi kata kunci yang lebih spesifik. Pengurangan bobot *diversity* memungkinkan model mempertahankan pemisahan leksikal tanpa mengorbankan koherensi internal topik.

4.5.5 Perbandingan Hasil Optuna

Perbandingan kinerja antar percobaan *hyperparameter tuning* dirangkum pada Tabel 4.11. Tabel ini menyajikan ringkasan nilai *coherence, diversity,* dan jumlah topik yang dihasilkan oleh masing-masing formulasi fungsi objektif.

Tabel 4.11 Perbandingan Hasil Optuna Tiap Percobaan

Rumus Skor	***Coherence***	***Diversity***	**Jumlah Topik**
coherence	0.478	-	4
coherence - penalty	0.508	-	6
coherence + diversity - penalty	0.514	0.766	12
coherence + 1/2diversity - penalty	0.524	0.773	15

Pada percobaan pertama, fungsi objektif yang hanya mengandalkan *coherence* menghasilkan nilai *coherence* ≈ 0,478 dengan empat topik. Meskipun nilai ini menunjukkan keterkaitan kata yang cukup baik, jumlah topik yang sangat terbatas dan dominasi satu topik

besar menunjukkan bahwa optimasi berbasis *coherence* saja belum mampu menghasilkan struktuk topik yang optimal.

Percobaan kedua menambahkan komponen penalti jumlah topik, yang menghasilkan peningkatan nilai *coherence* menjadi 0,508 dan jumlah topik meningkat menjadi enam. Hasil ini menunjukkan bahwa pengendalian jumlah topik efektif dalam memperbaiki distribusi klaster. Namun, karena fungsi objektif masih belum memasukkan *diversity*, keberagaman representasi kata antar topik masih relatif rendah.

Percobaan ketiga memperluas fungsi objektif dengan menambahkan komponen *diversity*, menghasilkan kombinasi nilai *coherence* ≈ 0,514 dan *diversity* ≈ 0,766, dengan dua belas topik. Hasil ini menunjukkan peningkatan yang jelas dalam keberagaman antar topik dibandingkan dua percobaan sebelumnya. Namun, sebagaimana ditunjukkan oleh distribusi nilai pada percobaan ketiga, kontribusi *diversity* terhadap skor akhir cenderung dominan, sehingga berpotensi mendorong konfigurasi dengan jumlah topik yang lebih besar namun stabilitas semantik yang belum sepenuhnya optimal.

Percobaan keempat melakukan penyesuaian lanjutan dengan mengurangi bobot komponen *diversity* menjadi setengah dari nilai aslinya. Konfigurasi terbaik pada percobaan ini menghasilkan nilai *coherence* tertinggi sebesar 0,524, *diversity* sebesar 0,773, dan lima belas topik. Dibandingkan percobaan ketiga, konfigurasi ini menunjukkan keseimbangan yang lebih baik antartopik, keberagaman, dan jumlah topik yang memadai tanpa dominasi klaster tertentu.

Secara keseluruhan, perbandingan pada Tabel 4.11 menunjukkan bahwa percobaan keempat dengan nilai masing-masing parameter `n_neighbors` = 14, `n_components` = 20, `min_dist` = 0,812, `min_cluster_size` = 324, dan `min_samples` = 11. memberikan hasil paling seimbang di antara seluruh percobaan. Formulasi fungsi objektif pada percobaan ini tidak hanya meningkatkan nilai *coherence*, tetapi juga mempertahankan tingkat *diversity* yang tinggi dengan struktur topik yang lebih stabil. Oleh karena itu, konfigurasi pada percobaan keempat ditetapkan sebagai konfigurasi final pemodelan topik dan digunakan pada tahap analisis struktur topik dan interpretasi tematik selanjutnya.

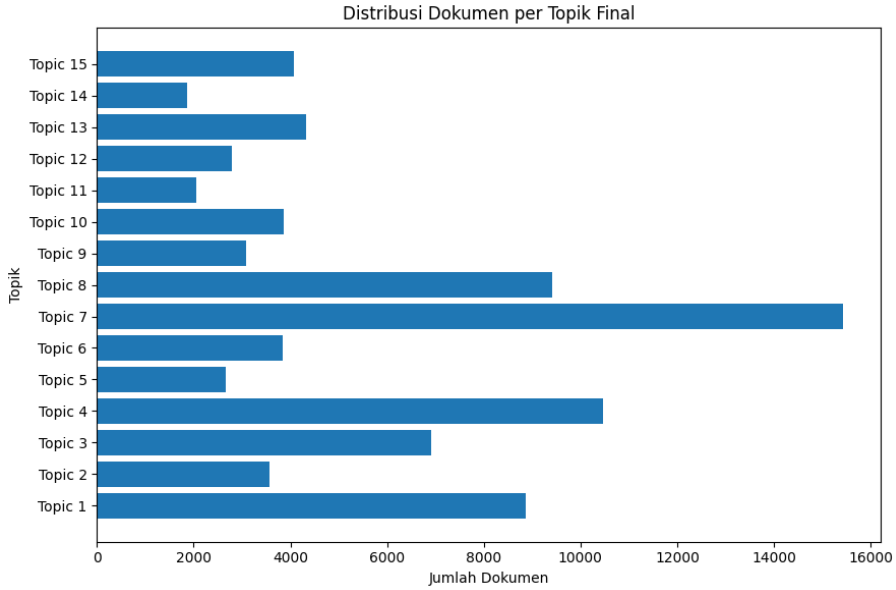
4.6 Struktur Topik Akhir Hasil Pemodelan

Berdasarkan konfigurasi final hasil hyperparameter tuning, model menghasilkan 15 topik yang dapat merepresentasikan struktur wacana publik. Bagian ini menyajikan struktur topik akhir dari sisi kuantitatif dan deskriptif, meliputi distribusi topik dan ukuran klaster,

representasi kata kunci berbobot, serta hubungan antartopik dalam ruang semantik. Pembahasan difokuskan pada validasi struktur hasil pemodelan, tanpa memasuki analisis interpretasi.

4.6.1 Distribusi Topik dan Ukuran Klaster

Distribusi jumlah dokumen pada masing-masing topik divisualisasikan pada Gambar 4.12. Visualisasi tersebut menampilkan jumlah dokumen per topik dalam bentuk diagram batang horizontal, sehingga perbedaan ukuran klaster antar topik dapat diamati secara langsung.



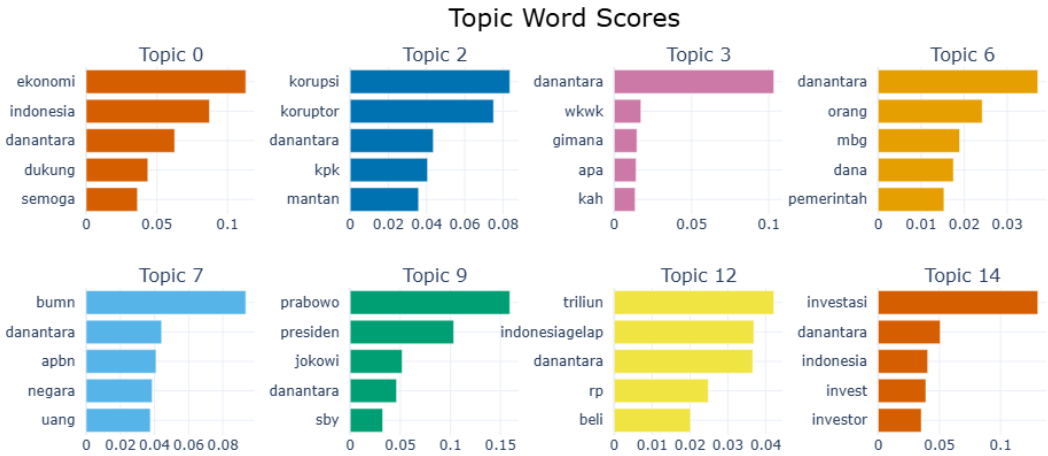
Gambar 4.12 Distribusi Dokumen per Topik Final

Terlihat bahwa ukuran klaster bervariasi secara bertahap, mulai dari topik dengan jumlah dokumen paling besar hingga topik dengan ukuran klaster yang lebih kecil. Terdapat satu topik dengan jumlah dokumen paling tinggi sebagai klaster dominan, diikuti oleh beberapa topik besar dan menengah dengan jumlah dokumen yang masih signifikan dan tidak terlalu jauh dari klaster dominan. Di sisi lain, topik-topik dengan ukuran klaster lebih kecil tetap memiliki jumlah dokumen yang memadai dan tidak tereduksi ke skala yang sangat rendah.

Keberadaan beberapa topik besar yang diimbangi oleh topik menengah dan kecil mengindikasikan bahwa konfigurasi final berhasil menjaga keseimbangan antartopik. Topik berukuran besar tidak mengaburkan keberadaan topik lain, sementara topik berukuran kecil jumlahnya relatif alami yang cukup untuk dianalisis. Secara keseluruhan, distribusi ukuran klaster pada Gambar 4.12 menunjukkan bahwa struktur topik hasil pemodelan relatif stabil.

4.6.2 Representasi Kata Kunci Berbobot Topik

Representasi internal masing-masing topik dianalisis melalui *Topic Word Scores* yang dihasilkan oleh mekanisme c-TF-IDF, yang merefleksikan kontribusi setiap kata terhadap pembentukan topik. Visualisasi *Topic Word Scores* tersebut disajikan dalam bentuk diagram batang pada Gambar 4.13 berikut ini.



Gambar 4.13 *Topic Word Scores Bar Chart*

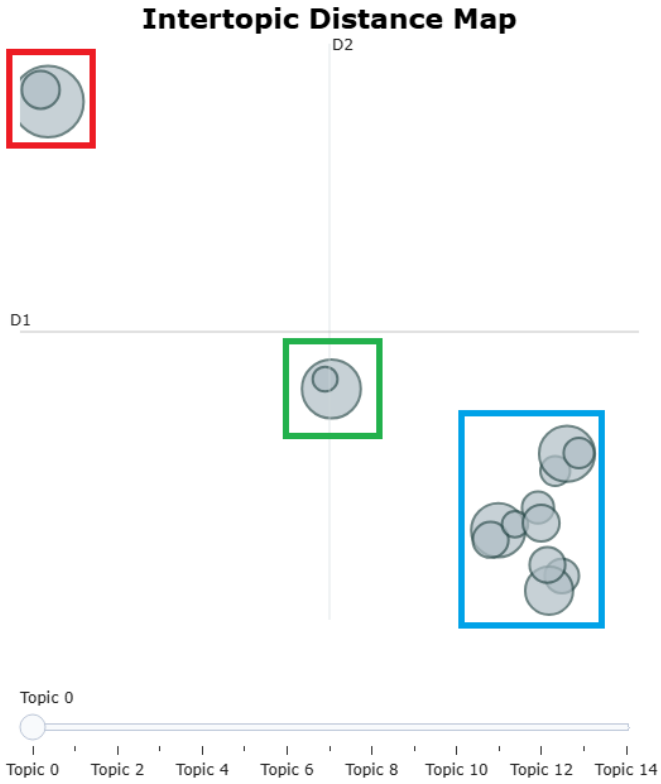
Setiap diagram batang menampilkan sejumlah kata dengan bobot tertinggi untuk satu topik, sehingga memperlihatkan struktur internal topik secara eksplisit. Visualisasi ini secara desain hanya menampilkan sebagian kata dan sebagian topik yang memiliki representasi kata paling kuat, dengan tujuan menyoroti karakter tematik utama secara ringkas dan informatif.

Pola pada Gambar 4.13 menunjukkan bahwa kata-kata dengan bobot tinggi dalam satu topik memiliki keterkaitan semantik yang konsisten dan saling mendukung seperti topik kedua yang berisi `korupsi` dan `koruptor`. Temuan ini mengindikasikan bahwa topik-topik yang divisualisasikan memiliki koherensi internal yang memadai.

Perbandingan antar diagram batang memperlihatkan bahwa meskipun token utama tertentu, seperti danantara atau indonesia, muncul lintas topik, kombinasi kata pendamping serta bobot relatifnya berbeda secara konsisten antar topik. Variasi pola bobot ini menunjukkan bahwa topik-topik berada pada ruang semantik yang terpisah. Pendekatan analisis berbasis *Topic Word Scores* berbobot ini dipilih karena lebih transparan dibandingkan visualisasi sebaran dokumen hasil reduksi dimensi.

4.6.3 *Intertopic Distance Map*

Intertopic Distance Map yang disajikan pada Gambar 4.14 merupakan proyeksi dua dimensi dari ruang *embedding* topik yang bertujuan untuk memvisualisasikan kedekatan semantik antartopik. Proyeksi ini dihasilkan melalui reduksi dimensi dari representasi vektor topik ke dalam dua sumbu abstrak, yang dilabeli sebagai D1 dan D2. Kedua sumbu tersebut berfungsi sebagai ruang proyeksi untuk memperlihatkan jarak relatif antartopik.



Gambar 4.14 *Intertopic Distance Map*

Ukuran dan posisi gelembung pada Gambar 4.14 merepresentasikan dua aspek utama, yaitu ukuran relatif topik dan kedekatan semantik antartopik. Topik-topik yang terletak berdekatan menunjukkan tingkat kemiripan semantik yang lebih tinggi, sedangkan topik yang terpisah jauh menunjukkan perbedaan semantik yang lebih kuat. Visualisasi tersebut memperlihatkan terbentuknya tiga kelompok utama topik:

a. Klaster Diskursus Ekonomi dan Kebijakan Institusional (Klaster Biru)

   Klaster biru mencakup topik 0, 1, 2, 4, 5, 7, 8, 9, 10, 11, dan 14, yang secara geometris saling berdekatan dalam ruang semantik. Kedekatan posisi antartopik menunjukkan tingkat kemiripan semantik yang tinggi, mengindikasikan bahwa wacana publik pada klaster ini terpusat pada domain ekonomi makro, kebijakan industri, investasi, fiskal,

energi, serta aktivitas korporasi negara. Kepadatan topik dalam satu wilayah ini menunjukkan bahwa diskursus ekonomi membentuk inti percakapan publik mengenai Danantara. Dominasi klaster ini dalam *Intertopic Distance Map* mengindikasikan bahwa narasi ekonomi dan kebijakan institusional berfungsi sebagai kerangka utama dalam struktur wacana secara keseluruhan.

b. Klaster Percakapan Informal (Klaster Hijau)

Klaster hijau yang terdiri dari topik 3 dan 13, menempati posisi yang terpisah dari klaster utama dan menunjukkan jarak semantik yang jelas terhadap klaster ekonomi. Topik-topik dalam klaster ini dicirikan oleh penggunaan bahasa informal, ungkapan spontan, humor, serta pertanyaan sederhana yang tidak berorientasi pada pembahasan serius mengenai kebijakan. Pemisahan posisi klaster ini menunjukkan bahwa model mampu membedakan fungsi komunikasi dalam wacana publik, khususnya antara konten informasional dan konten ekspresif. Keberadaan klaster ini menegaskan bahwa tidak seluruh percakapan publik berfungsi sebagai arena evaluasi kebijakan, melainkan juga sebagai ruang interaksi sosial dan ekspresi ringan.

c. Klaster Kritik dan Ketidakpercayaan terhadap Institusi (Klaster Merah)

Klaster merah yang mencakup topik 6 dan 12, berada pada posisi paling jauh dari klaster lainnya dalam ruang semantik. Jarak geometris yang lebih jauh dibandingkan jarak klister biru dan hijau menunjukkan perbedaan semantik yang signifikan. Topik-topik dalam klaster ini memperlihatkan orientasi wacana yang kritis dan konfrontatif terhadap pemerintah yang khususnya Danantara, ditandai oleh narasi kecurigaan, delegitimasi, serta ekspresi ketidakpercayaan yang kuat.

Secara keseluruhan, peta jarak antartopik pada Gambar 4.14 menunjukkan bahwa struktur topik hasil pemodelan memiliki pemisahan semantik yang jelas, baik dalam hal domain isu maupun karakter wacana. Visualisasi ini berfungsi sebagai validasi tambahan bahwa topik-topik yang terbentuk tidak hanya koheren secara internal, tetapi juga terorganisasi secara sistematis dalam ruang semantik.

4.7 Interpretasi Topik

Interpretasi topik pada bagian ini awalnya dilakukan secara mandiri oleh peneliti, dengan mengacu pada kata kunci berbobot berbasis *class-based* TF-IDF (c-TF-IDF) dan dokumen representatif (*representative documents*) hasil model BERTopic. Pendekatan ini dilakukan

untuk memenuhi evaluasi secara kualitatif dan membangun pemahaman awal mengenai makna setiap topik serta karakteristik wacana publik sebelum divalidasi oleh pihak lain.

Dokumen representatif yang digunakan dalam analisis ini bersumber dari dua mekanisme. Pertama, dokumen yang diidentifikasi secara otomatis oleh fungsi `get_topic_info()` pada BERTopic, yang memilih cuitan dengan probabilitas keanggotaan tertinggi terhadap suatu topik. Kedua, dilakukan penambahan dokumen secara manual oleh peneliti, dengan kriteria probabilitas keanggotaan topik di atas 0,9. Kombinasi ini bertujuan meningkatkan validitas interpretasi dengan tetap menjaga konsistensi terhadap struktur topik yang dihasilkan model. Ringkasan dokumen representatif untuk masing-masing topik disajikan pada Tabel 4.12.

Tabel 4.12 Dokumen Representatif

Topik Ke-	Dokumen Representatif
-1	• วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ วันนี้วันพระ https://t.co/h1KqQXHqrf • Quests > Airdrops
0	• Dukung Danantara, langkah Indonesia maju. • Yuk, dukung Danantara ekonomi Indonesia maju sejahtera. • Dukung Danantara ekonomi Indonesia maju. • salut sama Danantara, semoga tambah jaya. • danantara is here, siap-siap lihat karya hebat
1	• korupsi pertamina pengalihan danantara? • Danantara Pertamax. • Pertamina danantara. • Belum sempet bikin konten kasus Danantara dah muncul kasus Pertamina. Padet banget ya kasus tahun ini. • @hrdbacot Setiap ada isu besar seperti #indonesiagelap n DANANTARA muncul lah "barang baru" BBM oplosan, mmg kegiatan ngoplosnya baru kemaren ya, kok baru skrg ketangkepnya ?
2	• Danantara = Dana KORUPTOR. • Danantara investasi KORUPTOR. • 1 danantara ……koruptor.
3	• DANANTARA? • Danantara? 🤪 • Danantara 💩🔥🤣😅🤣 • How about danantara nantinya (?) • Danantara itu berperan gak min? • @pesolosep Mau nge jokes soal Danantara aja sih wkwkwkwk
4	• Danantara optimalkan aset BUMN investasi sektor dampak positif berkelanjutan. • Optimalkan aset BUMN investasi sektor berkelanjutan Danantara, dampak positif.

	• HIPMI Nilai Danantara Wujud Komitmen Pemerintah dalam Reformasi BUMN @BPPHIPMI @KemenBUMN • DANANTARA menjadi inovasi dalam pengelolaan aset negara, agar lebih produktif dan berdaya saing optimal bagi perekonomian nasional. • Danantara optimalkan aset BUMN sektor investasi berdampak positif Berkelanjutan
5	• Presiden Republik Indonesia, Prabowo Subianto, resmi meluncurkan Badan • Pengelola Investasi (BPI) Daya Anagata Nusantara (Danantara) Istana KepresidenanSenin (24/2/2025). • Presiden Prabowo Subianto resmi meluncurkan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara) Senin (24/2). • Presiden Prabowo Subianto resmi meresmikan Badan Pengelola Investasi Daya Anagata Nusantara (BPI Danantara) Istana Negara, Jakarta, Senin (24/2/2025).
6	• Program 3 Juta Rumah Presiden Prabowo cepat terealisasi Danantara. • Pasal 3 UU Keuangan Negara: penerimaan negara masuk APBN, surplus APBN modal BUMN persetujuan DPR. Inpres No. 1 Thn 2025 melanggar Pasal 3 UUKeuangan Negara, batal hukum. Modal Danantara = hasil korupsi. • 3. Danantara. • @kumparan Prabowo bikin Danantara bikin bank emas bikin koprasi desa, bukan untuk memakmurkan rakyat nya, untuk memakmurkan para koruptor • @CNNIndonesia Ini mengarah ke prdiksi gw sebelumnya, sbelum akhir 2025, kurs rupiah/dolar akn tmbus Rp17.000. Pasar tdk sudi ikut pemerintah. Danantara akn membuat investor asing cemas utk parkir dana di Indons krn lihat riwayat bbrp proyek besar pemerintah pada jeblok. 😳
7	• Danantara BUMN? • 8 bumn danantara? • BUMN -> DANANTARA. • Sedang membaca² danantara 👀 Berarti kementerian BUMN nnt fungsinya apa ? 👀 • @realfedinuril @Uki23 Nah ini,apakah APBN, atau dari danantara itu sendiri?
8	• Transparansi anggaran Danantara efisiensi. • Investasi efisiensi anggaran transparansi Danantara. • Danantara, transparansi investasi efisiensi anggaran. • @RahmahSeikha Kuatkan fondasi Danantara dengan anggaran yang efisien dan transparan • Anggaran lebih efisien, transparansi terjaga berkat Danantara #LangkahTepatEkonomiKuat https://t.co/BQvObXj0yi
9	• Presiden Prabowo luncurkan danantara. • DANANTARA PRABOWO. • DANANTARA PRABOWO. • PANGLIMA TNI DAMPINGI PRESIDEN RI RESMIKAN DANANTARA https://t.co/Y7LG7NDJFX via @TABLOID NUSANTARA • Danantara ini akal² pemerintahan di bwh presiden gemoy @prabowo @Gerindra Melanjutkan rencana eks presiden incapable @jokowi lewat @erickthohir bendaharawan pribadinya di @KemenBUMN untuk bisa njarah dan ngerampok kekayaan negara dgn mulus. Terpusat pada satu institusi lbh aman.
10	• Danantara dorong investasi transisi energi. • Danantara Dorong Investasi Transisi Energi.

	• Danantara Mendorong Investasi Transisi Energi Terbarukan. • Danantara, harapan baru bagi para pelaku industri kreatif digital untuk go global dan mendunia. #IndustriKreatifDigital #Danantara • "Pemerintah Kota Yogyakarta tentu siap berkolaborasi dan bergabung dalam proyek Danantara. Selain mengurangi timbunan sampah, program ini juga bisa menghasilkan manfaat baru berupa energi listrik," ujar Hasto Wardoyo.
11	• BCA masuk Danantara. • bca masuk danantara? • danantara, BCA. • sebener e ini dividen e kan yaa buat danantara, cuma takut worst case tbtb uang ilang, mau ganti ke bca tapi udah kebanyakan bank yg kepake 🫣 • @AtanCrowford @Just_MasBima @irwndfrry Itu baca dimana bang BCA masuk danantara? (Bertanya dengan nada santai) https://t.co/uqic14WW6F
12	• Danantara ndasmu! IndonesiaGelap2025. • Danantara IndonesiaGelap2025. • IndonesiaGelap2025 Danantara. • Danantara target 75o T dalam 5 tahun ????? Semua bumn di kelolah oleh danatara kalau 750 dalam 5 tahun berarti 1 tahun target nya 150 T dividen nya . Aneh cuma 150 T dividen nya seharusnya target duviden itu 500 T dalam 1 tahun dividennya . Emmmm abal abak • @axelovers125 reminder, kalik aja lupa belum ada setahun si ndut -banyak ASN di cutoff + efisiensi demi MBG -APBN membengkak demi MBG -indonesia gelap, RUU TNI -jurnalis dapet ancaman kepala babi -danantara, dipimpin sama koruptor -banyak siswa keracunan MBG -demo agustus 2025 7 sipil MD
13	• Tolong jelasin bahasa bayi danantara. • Jelasin danantara bahasa bayi. • danantara jelasin bahasa bayi? • @tanyakanrl Jelasin tentang Danantara ini pake bahasa unang dong • Perlu banyak tenaga untuk baca baca soal danantara dengan bahasa bayi, kl lg cape otak gue lemot bgt 🤯
14	• Ekonomi investasi Danantara • investasi Danantara • Danantara investasi • Dengan hadirnya Danantara, peluang investasi semakin terbuka lebar • pasti bisa, nanti ada danantara mungkin mau investasi disini 😋 https://t.co/J1dEOk1rCS

Analisis secara kualitatif terhadap Tabel 4.13 menunjukkan bahwa dokumen representatif memiliki korespondensi yang kuat dengan kata kunci dominan pada tiap topik. Topik seperti dukungan terhadap Danantara, diwakili oleh cuitan singkat dan repetitif yang menegaskan optimisme ekonomi. Sebaliknya, topik yang bersifat kritis atau oposisi ditandai oleh dokumen dengan bahasa konfrontatif, penggunaan istilah hukum, atau ekspresi emosional yang eksplisit.

Sementara itu, topik percakapan informal dan kebingungan publik memuat cuitan dengan gaya tutur kasual, humor, atau pertanyaan langsung, yang menunjukkan perbedaan fungsi komunikasi dalam wacana publik.

Untuk merangkum hasil pemaknaan tersebut secara sistematis, Tabel 4.13 menyajikan interpretasi mandiri akhir setiap topik berdasarkan sintesis antara kata kunci c-TF-IDF dan isi dokumen representatif. Tabel ini berfungsi sebagai analisis kualitatif yang ditopang oleh bukti tekstual dari representasi dokumen.

Tabel 4.13 Interpretasi Topik Mandiri

No	**Kata Kunci Utama**	**Interpretasi**
-1	['นน', 'นพระ', 'x648', 'x624', 'та', 'babik', 'goat', 'emyu', 'acakadyt', 'airdrops']	outlier
0	['ekonomi', 'indonesia', 'danantara', 'dukung', 'semoga', 'pertumbuhan', 'maju', 'langkah', 'rakyat', 'nasional']	Narasi dukungan dan optimisme ekonomi
1	['pertamina', 'pertamax', 'isu', 'danantara', 'pengalihan', 'oplosan', 'indonesiagelap', 'korupsi', 'harga', 'pertalite']	Isu BBM, Pertamina, dan kecurigaan pengalihan isu
2	['korupsi', 'koruptor', 'danantara', 'kpk', 'mantan', 'bpk', 'hukum', 'audit', 'pengawas', 'diaudit']	Tuduhan korupsi dan delegitimasi Danantara
3	['danantara', 'wkwk', 'gimana', 'apa', 'kah', 'gini', 'wkwkwk', 'pak', 'eh', 'pas']	Percakapan informal dan respons bercanda
4	['aset', 'berkelanjutan', 'positif', 'bumn', 'investasi', 'negara', 'sektor', 'danantara', 'optimalisasi', 'optimalkan']	Optimalisasi aset BUMN dan investasi berkelanjutan
5	['badan', 'pengelola', 'nusantara', 'anagata', 'investasi', 'daya', 'bpi', '2025', 'presiden', 'prabowo']	Peluncuran resmi Danantara oleh Presiden
6	['danantara', 'orang', 'mbg', 'dana', 'pemerintah', 'rakyat', 'asing', 'negara', 'saham', 'uu']	Kontroversi legalitas, dana publik, dan kepemilikan asing
7	['bumn', 'danantara', 'apbn', 'negara', 'uang', 'masuk', 'dividen', 'rakyat', 'duit', 'dana']	Arsitektur pendanaan negara melalui BUMN dan APBN
8	['efisiensi', 'anggaran', 'transparansi', 'danantara', 'mbg', 'transparan', 'investasi', 'hasil', 'dana', 'efisien']	Transparansi anggaran dan efisiensi investasi
9	['prabowo', 'presiden', 'jokowi', 'danantara', 'sby', 'luncurkan', 'indonesia', 'presidenprabowo', 'tepatsolutifindonesia', 'berita']	Pemberitaan politik terkait Prabowo dan peluncuran Danantara
10	['energi', 'kerja', 'hilirisasi', 'industri', 'lapangan', 'proyek', 'danantara', 'transisi', 'investasi', 'sektor']	Investasi hilirisasi industri dan transisi energi

11	['bank', 'bca', 'mandiri', 'pindah', 'danantara', 'swasta', 'duit', 'bri', 'uang', 'tabungan']	Keterlibatan sektor perbankan dalam Danantara
12	['triliun', 'indonesiagelap', 'danantara', 'rp', 'beli', 'ndasmu', 'rakyat', 'indonesiagelap2025', 'negara', 'dana']	Narasi oposisi ekstrem dan kritik keras
13	['bahasa', 'bang', 'paham', 'jelasin', 'bahas', 'baca', 'danantara', 'bayi', 'penjelasan', 'tolong']	Permintaan penjelasan dan kebingungan publik
14	['investasi', 'danantara', 'indonesia', 'invest', 'investor', 'global', 'ekonomi', 'strategis', 'nasional', 'sektor']	Investasi strategis dan positioning ekonomi nasional

Setelah proses interpretasi mandiri dilakukan, peneliti melibatkan *human annotator* untuk menilai kesesuaian antara kata kunci, dokumen representatif, serta koherensi makna setiap topik yang dihasilkan. Dengan demikian, konsistensi antara keluaran model, bukti empiris, dan batasan inferensi yang dapat ditarik dari data media sosial tetap terjaga.

Sebanyak dua orang *human annotator* berpartisipasi dalam proses validasi interpretasi topik. Kedua *annotator* memiliki latar belakang keilmuan yang berbeda guna memberikan sudut pandang yang saling melengkapi, yakni bidang hubungan internasional dan linguistik.

Annotator pertama, Deva Soka Faridh Aziz, merupakan mahasiswa Jurusan Hubungan Internasional dengan pengalaman dalam praktik diplomasi, termasuk sebagai *Best Delegate* dalam simulasi konferensi Organisasi Islam Dunia. Deva memberikan penilaian rendah (skor 2 dari 5) terhadap interpretasi Topik 2 dan Topik 6, yang menunjukkan ketidaksesuaian antara interpretasi peneliti dan bukti yang tersaji dalam data. Menurut Deva, interpretasi awal Topik 2 terlalu memfragmentasikan berbagai isu tanpa dukungan bukti konkret, sehingga lebih merepresentasikan dugaan dan kecurigaan publik daripada fakta yang terverifikasi. Sementara itu, interpretasi awal Topik 6 dinilai mencakup terlalu banyak variabel, seperti kepemilikan asing, korupsi, dan pelemahan nilai tukar, yang tidak seluruhnya didukung oleh data dalam kluster topik tersebut.

Berdasarkan masukan tersebut, interpretasi Topik 2 direvisi dari “Tuduhan korupsi dan delegitimasi Danantara” menjadi “Ekspresi Tuduhan dan Kecurigaan Publik terhadap Integritas Danantara”, sedangkan interpretasi Topik 6 direvisi dari “Kontroversi legalitas, dana publik, dan kepemilikan asing” menjadi “Perdebatan Publik mengenai Aspek Legal dan Kebijakan Ekonomi dalam Pembentukan Danantara”. Meskipun demikian, Deva menyatakan bahwa secara keseluruhan kumpulan topik yang dihasilkan tetap dapat mengarah pada analisis isu-isu tematik yang bermakna.

Annotator kedua, Ainun Safitri, merupakan lulusan magister linguistik Universitas Indonesia dan berpengalaman sebagai data *annotator*. Ainun memberikan penilaian positif terhadap seluruh interpretasi topik, dengan tidak adanya skor negatif dan seluruh penilaian berada pada rentang netral hingga positif (≥3). Ainun juga menyatakan bahwa struktur topik dan interpretasi yang dihasilkan telah relevan dengan tujuan penelitian.

Dengan mempertimbangkan penilaian dan masukan dari kedua *annotator*, interpretasi akhir topik disesuaikan dan diperkuat secara konseptual. Hasil interpretasi topik setelah melalui proses validasi human annotator disajikan pada Tabel 4.14.

Tabel 4.14 Interpretasi Pasca Validasi dari *Annotator*

No	**Kata Kunci Utama**	**Interpretasi**
-1	['นน', 'นพระ', 'x648', 'x624', 'та', 'babik', 'goat', 'emyu', 'acakadyt', 'airdrops']	outlier
0	['ekonomi', 'indonesia', 'danantara', 'dukung', 'semoga', 'pertumbuhan', 'maju', 'langkah', 'rakyat', 'nasional']	Narasi dukungan dan optimisme ekonomi
1	['pertamina', 'pertamax', 'isu', 'danantara', 'pengalihan', 'oplosan', 'indonesiagelap', 'korupsi', 'harga', 'pertalite']	Isu BBM, Pertamina, dan kecurigaan pengalihan isu
2	['korupsi', 'koruptor', 'danantara', 'kpk', 'mantan', 'bpk', 'hukum', 'audit', 'pengawas', 'diaudit']	Ekspresi Tuduhan dan Kecurigaan Publik terhadap Integritas Danantara
3	['danantara', 'wkwk', 'gimana', 'apa', 'kah', 'gini', 'wkwkwk', 'pak', 'eh', 'pas']	Percakapan informal dan respons bercanda
4	['aset', 'berkelanjutan', 'positif', 'bumn', 'investasi', 'negara', 'sektor', 'danantara', 'optimalisasi', 'optimalkan']	Optimalisasi aset BUMN dan investasi berkelanjutan
5	['badan', 'pengelola', 'nusantara', 'anagata', 'investasi', 'daya', 'bpi', '2025', 'presiden', 'prabowo']	Peluncuran resmi Danantara oleh Presiden
6	['danantara', 'orang', 'mbg', 'dana', 'pemerintah', 'rakyat', 'asing', 'negara', 'saham', 'uu']	Perdebatan Publik mengenai Aspek Legal dan Kebijakan Ekonomi dalam Pembentukan Danantara
7	['bumn', 'danantara', 'apbn', 'negara', 'uang', 'masuk', 'dividen', 'rakyat', 'duit', 'dana']	Arsitektur pendanaan negara melalui BUMN dan APBN
8	['efisiensi', 'anggaran', 'transparansi', 'danantara', 'mbg', 'transparan', 'investasi', 'hasil', 'dana', 'efisien']	Transparansi anggaran dan efisiensi investasi

9	['prabowo', 'presiden', 'jokowi', 'danantara', 'sby', 'luncurkan', 'indonesia', 'presidenprabowo', 'tepatsolutifindonesia', 'berita']	Pemberitaan politik terkait Prabowo dan peluncuran Danantara
10	['energi', 'kerja', 'hilirisasi', 'industri', 'lapangan', 'proyek', 'danantara', 'transisi', 'investasi', 'sektor']	Investasi hilirisasi industri dan transisi energi
11	['bank', 'bca', 'mandiri', 'pindah', 'danantara', 'swasta', 'duit', 'bri', 'uang', 'tabungan']	Keterlibatan sektor perbankan dalam Danantara
12	['triliun', 'indonesiagelap', 'danantara', 'rp', 'beli', 'ndasmu', 'rakyat', 'indonesiagelap2025', 'negara', 'dana']	Narasi oposisi ekstrem dan kritik keras
13	['bahasa', 'bang', 'paham', 'jelasin', 'bahas', 'baca', 'danantara', 'bayi', 'penjelasan', 'tolong']	Permintaan penjelasan dan kebingungan publik
14	['investasi', 'danantara', 'indonesia', 'invest', 'investor', 'global', 'ekonomi', 'strategis', 'nasional', 'sektor']	Investasi strategis dan positioning ekonomi nasional

Secara keseluruhan, struktur topik yang terbentuk menunjukkan bahwa percakapan publik mengenai Danantara tergabung ke dalam beberapa domain diskursus utama. Domain tersebut mencakup narasi dukungan dan optimisme ekonomi, kekhawatiran terhadap tata kelola dan korupsi, dinamika politik dan legitimasi kekuasaan, Isu teknis fiskal dan investasi, serta ekspresi kebingungan publik dan percakapan informal. Keragaman ini menegaskan bahwa Danantara tidak dipersepsikan secara tunggal sebagai instrumen ekonomi, melainkan sebagai entitas yang berada pada irisan hukum, politik, dan kepercayaan publik.

Bagian selanjutnya menguraikan masing-masing domain diskursus tersebut secara lebih rinci melalui pengelompokan topik-topik yang memiliki kedekatan tematik, dengan tetap merujuk pada kata kunci dominan dan dokumen representatif sebagai dasar interpretasi.

4.7.1 Klaster *Outlier* dan Konten NonRelevan

Topik -1 diklasifikasikan sebagai klaster *outlier* karena didominasi oleh token acak, lintas bahasa, serta istilah yang tidak memiliki keterkaitan semantik dengan konteks Danantara. Kata kunci seperti นน, x648, x624, maupun token asing lainnya tidak membentuk struktur yang bermakna.

Dokumen representatif pada klaster ini memperkuat karakteristik tersebut. Cuitan-cuitan yang muncul sebagian besar berupa pengulangan frasa tidak bermakna, fragmen teks lintas bahasa tanpa konteks, atau tautan yang sama sekali tidak berkaitan dengan Danantara. Pola ini

mengindikasikan keberadaan *noise* yang tidak sepenuhnya dapat dieliminasi melalui pembersihan berbasis duplikasi dan anomali temporal.

Keberadaan klaster *outlier* ini secara metodologis menunjukkan bahwa model masih mampu mengisolasi konten yang tidak relevan ke dalam satu klaster tersendiri, sehingga tidak mengganggu struktur topik lainnya. Oleh karena itu, Topik -1 secara sengaja dikeluarkan dari analisis tematik lanjutan karena tidak merepresentasikan wacana publik yang bermakna dalam konteks pembahasan Danantara.

4.7.2 Narasi Dukungan dan Optimisme Ekonomi

Topik 0 dan Topik 14 merepresentasikan narasi dukungan dan optimisme ekonomi terhadap Danantara. Kata kunci dominan seperti `ekonomi`, `pertumbuhan`, `investasi`, `strategis`, dan `nasional` menempatkan Danantara sebagai instrumen penguatan ekonomi jangka panjang serta simbol harapan terhadap kemajuan ekonomi Indonesia.

Dokumen representatif pada kedua topik ini menunjukkan pola ekspresi yang sangat afirmatif dan deklaratif. Karena sangat deklaratif, kedua topik ini membuat kecurigaan bahwa ekspresi ini dibuat secara tidak alami. Pada Topik 0, cuitan-cuitan seperti “Dukung Danantara, langkah Indonesia maju” dan “Yuk, dukung Danantara ekonomi Indonesia maju sejahtera” menampilkan bentuk dukungan eksplisit yang sangat repetitif tanpa adanya konteks yang kritis. Sementara itu, Topik 14 memuat cuitan yang menekankan aspek investasi, seperti “Danantara investasi” dan “Dengan hadirnya Danantara, peluang investasi semakin terbuka lebar”, yang mencerminkan ekspektasi terhadap peran lembaga ini dalam menarik modal dan memperkuat posisi ekonomi nasional. Namun, sifatnya sama dengan topik 0 yang sangat repetitif.

Karakteristik dokumen representatif yang relatif singkat, berulang, dan minim argumentasi teknis menunjukkan bahwa dukungan pada klaster ini lebih bersifat normatif dibandingkan analitis. Publik dalam klaster ini tidak secara mendalam membahas mekanisme kebijakan atau implikasi struktural Danantara, melainkan mengekspresikan sikap positif secara langsung dan kolektif.

4.7.3 Kekhawatiran Terhadap Tata Kelola dan Korupsi

Topik 2, 8, dan 12 membentuk domain diskursus yang berfokus pada kritik, kecurigaan, ketidakpercayaan, dan tuntutan akuntabilitas terhadap Danantara. Kesamaan ketiganya terletak pada orientasi skeptisisme terhadap pemerintahan, namun masing-masing topik menunjukkan

karakter wacana yang berbeda, baik dari sisi kata kunci dominan maupun pola isi dokumen representatif.

Topik 2 merepresentasikan kritik yang bersifat delegitimatif dan berbasis tuduhan langsung. Kata kunci seperti `korupsi`, `koruptor`, `KPK`, `BPK`, `audit`, dan `diaudit` menunjukkan fokus kuat pada isu hukum dan pengawasan. Dokumen representatif pada topik ini, seperti "Danantara = Dana KORUPTOR" atau "Danantara investasi KORUPTOR", mengindikasikan *framing* Danantara sebagai entitas yang sejak awal dicurigai bermasalah secara moral dan hukum, tanpa melalui argumentasi prosedural yang rinci. Dengan demikian, Topik 2 mencerminkan bentuk kritik yang mengupayakan bentuk delegitimasi terhadap institusi.

Berbeda dengan Topik 2, Topik 8 menampilkan kritik yang lebih teknokratis dan berorientasi tata kelola. Kata kunci seperti `transparansi`, `efisiensi`, `anggaran`, dan `hasil` menempatkan Danantara dalam kerangka evaluasi kinerja dan akuntabilitas pengelolaan dana publik. Dokumen representatif pada topik ini, seperti "Transparansi anggaran Danantara efisiensi" atau "Investasi efisiensi anggaran transparansi Danantara", menunjukkan tuntutan terhadap prinsip *good governance* (Rahmanto, 2022). Kritik pada Topik 8 tidak secara eksplisit menuduh penyimpangan, tetapi menekankan kebutuhan akan kejelasan mekanisme, keterbukaan informasi berupa transparansi, dan rasionalitas pengelolaan anggaran.

Sementara itu, Topik 12 merepresentasikan bentuk kritik yang lebih ekstrem dan emosional. Kata kunci seperti `triliun`, `IndonesiaGelap`, dan `ndasmu` menunjukkan intensifikasi ekspresi ketidakpercayaan. Dokumen representatif seperti "Danantara ndasmu! IndonesiaGelap2025" atau "IndonesiaGelap2025 Danantara" menampilkan gaya bahasa yang kasar dan hiperbolik. Kritik pada topik ini diarahkan pada aspek delegitimasi politik yang lebih luas, sering kali dikaitkan dengan sentimen pesimistis terhadap kondisi negara secara keseluruhan.

Keberadaan tiga topik yang terpisah menunjukkan kritik yang bergerak dari tuduhan hukum, tuntutan tata kelola yang rasional, hingga ekspresi emosional. Diferensiasi ini menegaskan bahwa wacana negatif dipahami sebagai kumpulan posisi yang memiliki dasar, gaya, dan intensitas yang berbeda dalam merespons keberadaan Danantara.

4.7.4 Dinamika Politik dan Legitimasi Kekuasaan

Topik 5 dan Topik 9 menempatkan Danantara secara eksplisit dalam konteks politik negara dan tokoh kekuasaan. Kata kunci seperti `presiden`, `Prabowo`, `Jokowi`, `peluncuran`,

dan `berita` menunjukkan bahwa pembentukan Danantara dipersepsikan publik sebagai peristiwa politik melibatkan penguasa.

Dokumen representatif pada Topik 5 didominasi oleh cuitan informatif seperti peresmian Danantara oleh Presiden Prabowo Subianto. Cuitan seperti "Presiden Republik Indonesia, Prabowo Subianto, resmi meluncurkan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara)" menunjukkan pola komunikasi yang berfungsi menyebarkan legitimasi formal lembaga. Narasi pada topik ini memberi penekanan pada waktu, tempat, dan otoritas peresmian, tanpa disertai evaluasi kritis terhadap kebijakan tersebut.

Topik 9 mempunyai gaya bahasa yang lebih sloganistik dan repetitif. Dokumen representatif seperti "DANANTARA PRABOWO" atau "Presiden Prabowo luncurkan Danantara" menunjukkan personalisasi lembaga pada figur presiden. Selain itu, kehadiran cuitan yang menyinggung figur lain, seperti Panglima TNI, memperluas *framing* Danantara sebagai proyek negara yang melibatkan struktur kekuasaan yang lebih luas. Hal ini sesuai dengan latar belakang Presiden Prabowo sebagai mantan Letnan Dua TNI Angkatan Darat. Pola ini menandakan bahwa legitimasi Danantara dalam wacana publik sebagian dibangun melalui asosiasi langsung dengan simbol pertahanan negara.

Dominasi aktor politik dalam kedua topik ini menunjukkan bahwa persepsi publik terhadap Danantara masih sangat bergantung pada sumber legitimasi politik. Lembaga ini dipahami sebagai ekstensi dari kekuasaan eksekutif, sehingga penerimaan maupun penolakannya berpotensi mengikuti dinamika kepercayaan terhadap figur politik yang melekat padanya. Dengan demikian, Topik 5 dan 9 merefleksikan bagaimana kebijakan ekonomi negara tidak pernah sepenuhnya terlepas dari arena politik.

4.7.5 Isu Teknis Fiskal dan Investasi

Topik 4, 7, 10, dan 11 membentuk klaster diskursus yang berfokus pada aspek teknis dan operasional Danantara, khususnya terkait pengelolaan aset BUMN, arsitektur pendanaan negara, investasi sektor strategis, serta keterlibatan aktor keuangan seperti perbankan. Kata kunci dominan pada topik-topik ini menunjukkan orientasi wacana pada mekanisme, sektor, dan implikasi ekonomi konkret.

Topik 4 merepresentasikan narasi mengenai optimalisasi aset BUMN dan investasi berkelanjutan. Dokumen representatif pada topik ini, seperti "Danantara optimalkan aset BUMN investasi sektor berdampak positif berkelanjutan", menampilkan pola bahasa yang relatif formal. Cuitan-cuitan tersebut mencerminkan pemaknaan Danantara untuk

meningkatkan nilai aset negara dan mendorong investasi yang berorientasi jangka panjang. Konsistensi antara kata kunci dan dokumen representatif menunjukkan bahwa topik ini memuat diskursus kebijakan pada tujuan dan manfaat ekonomi.

Topik 7 menyoroti diskursus mengenai hubungan antara Danantara, BUMN, dan APBN. Dokumen representatif seperti "Danantara BUMN?" atau "Berarti kementerian BUMN nanti fungsinya apa?" memperlihatkan pola pertanyaan dan eksplorasi konseptual. Berbeda dengan Topik 4 yang cenderung formal, Topik 7 mencerminkan upaya publik untuk memahami restrukturisasi kelembagaan dan implikasi fiskal dari pembentukan Danantara. Wacana pada topik ini menunjukkan adanya kebutuhan klarifikasi terkait peran BUMN dan distribusi kewenangan keuangan negara.

Topik 10 berfokus pada sektor strategis, khususnya energi, hilirisasi industri, dan transisi energi. Dokumen representatif seperti "Danantara mendorong investasi transisi energi terbarukan" menunjukkan konsistensi antara kata kunci dan narasi isi. Selain sektor energi, muncul pula cuitan yang mengaitkan Danantara dengan industri kreatif yang menandakan perluasan persepsi publik terhadap cakupan sektor yang dapat disentuh oleh Danantara. Topik ini mencerminkan ekspektasi publik terhadap peran Danantara dalam mendorong transformasi struktural ekonomi.

Topik 11 merepresentasikan diskursus mengenai keterlibatan sektor perbankan, khususnya bank-bank besar seperti BCA, Mandiri, dan BRI. Dokumen representatif pada topik ini memperlihatkan kombinasi antara pernyataan informatif dan kekhawatiran personal, seperti spekulasi mengenai pemindahan dana atau risiko kehilangan tabungan. Pola ini menunjukkan bahwa keterlibatan bank dalam Danantara berdampak pada persepsi publik mengenai risiko finansial di tingkat individu.

Secara keseluruhan, keempat topik ini menunjukkan adanya segmen publik yang terlibat dalam menilai Danantara melalui lensa teknis. Wacana yang muncul berupaya memahami bagaimana Danantara beroperasi, sektor apa yang disentuh, dan implikasi apa yang mungkin timbul bagi keuangan negara. Hal ini menunjukkan bahwa sebagian diskursus publik membahas Danantara secara substantif, meskipun tetap muncul bersamaan dengan narasi dukungan dan kritik politik.

4.7.6 Kebingungan Publik dan Percakapan Informal

Topik 3 dan Topik 13 membentuk domain wacana nonsubstantif yang berbeda secara fungsi, tetapi saling berkaitan dalam menunjukkan dinamika pemahaman publik terhadap

Danantara. Kedua topik ini berfokus pada respons spontan, gaya komunikasi informal, serta kebutuhan akan penjelasan dasar.

Topik 3 didominasi oleh ekspresi informal, humor, dan reaksi singkat. Dokumen representatif seperti "Danantara? 🤪", "Danantara 💩🔥🤣😅🤣", atau sekadar "DANANTARA?" menunjukkan pola respons reaktif yang bersifat ekspresif. Penggunaan emoji, tawa, dan tanda tanya berulang mengindikasikan bahwa Danantara hadir sebagai objek percakapan kasual. Wacana pada topik ini berfungsi lebih sebagai penanda keterlibatan sosial dan partisipasi ringan dalam arus percakapan publik, tanpa upaya eksplisit untuk memahami atau menilai substansi lembaga tersebut secara lebih mendalam.

Berbeda dengan Topik 3, Topik 13 merepresentasikan kebingungan konseptual yang lebih eksplisit. Dokumen representatif seperti "Tolong jelasin bahasa bayi danantara" atau "danantara jelasin bahasa bayi?" menunjukkan adanya permintaan klarifikasi yang sadar dan langsung. Gaya bahasa yang digunakan tetap informal, namun mengandung kebutuhan informasi. Hal ini mengindikasikan bahwa sebagian publik kurang memahami konsep dari Danantara dan pada akhirnya mengekspresikan kebutuhan akan penjelasan tersebut ke dalam bentuk yang lebih kasual dan informal.

Perbedaan karakter antara Topik 3 dan Topik 13 menunjukkan dua bentuk respons yang berbeda. Topik 3 mencerminkan reaksi emosional yang tidak berorientasi pada pencarian makna, sedangkan Topik 13 menunjukkan adanya permintaan penjelasan secara eksplisit oleh pengguna. Keberadaan kedua topik ini menegaskan bahwa tidak seluruh percakapan publik mengenai Danantara bergerak pada level evaluatif atau kritis. Sebagian wacana bersifat ringan dan menunjukkan keterbatasan pemahaman.

Secara analitis, klaster ini penting karena menunjukkan ruang komunikasi publik yang sering terabaikan dalam analisis kebijakan berbasis sentimen atau polaritas. Kebingungan dan percakapan informal menunjukkan indikator adanya tantangan dalam penyampaian informasi kebijakan. Jika tidak direspons, domain wacana ini berpotensi menjadi pintu masuk bagi misinterpretasi, spekulasi, dan penguatan narasi negatif di klaster lain. Dengan demikian, Topik 3 dan Topik 13 melengkapi pemahaman mengenai wacana publik dari diskursus serius hingga ekspresi sosial yang sederhana.

4.8 Diskusi dan Implikasi

Hasil pemodelan topik menggunakan BERTopic dengan optimalisasi UMAP dan HDBSCAN menunjukkan bahwa wacana publik mengenai Danantara tersusun dalam berbagai

domain diskursus yang mencakup optimisme ekonomi, pembahasan teknokratis, kritik terhadap tata kelola, politisasi kebijakan, hingga ekspresi kebingungan dan interaksi informal. Struktur ini menegaskan bahwa Danantara dipersepsikan sebagai entitas multidimensional yang berada pada persilangan kebijakan ekonomi, legitimasi politik, dan dinamika komunikasi media sosial.

Narasi dukungan dan optimisme ekonomi mencerminkan harapan publik terhadap peran Danantara dalam mendorong pertumbuhan ekonomi dan investasi strategis nasional. Namun, optimisme tersebut berjalan berdampingan dengan wacana kritis terkait isu korupsi, transparansi, dan legalitas, yang muncul dalam korpus data. Pola ini menunjukkan penerimaan terhadap tujuan kebijakan tidak selalu disertai dengan kepercayaan penuh terhadap mekanisme dan tata kelola institusionalnya.

Di sisi lain, keberadaan diskursus teknokratis mengenai pengelolaan aset BUMN, sektor strategis, dan keterlibatan perbankan menunjukkan adanya segmen publik yang berupaya memahami Danantara secara lebih mendalam. Bersamaan dengan itu, kemunculan topik kebingungan dan percakapan informal memperlihatkan bahwa partisipasi publik di media sosial tidak selalu bersifat formal. Ekspresi kebingungan, pertanyaan sederhana, dan humor mencerminkan gaya komunikasi digital yang cair.

Tidak seluruh isu yang beredar dalam ruang publik dapat tertangkap secara rinci sebagai topik yang berdiri sendiri dalam hasil pemodelan topik. Hal ini berkaitan dengan karakteristik distribusi bahasa alami di media sosial serta mekanisme pembentukan klaster pada model berbasis *embedding*. Isu-isu tertentu yang sempat menjadi perbincangan intens sering kali diekspresikan melalui variasi istilah yang tidak konsisten, frekuensi penggunaan kata kunci yang relatif rendah, atau muncul secara sporadis dalam konteks yang beragam. Akibatnya, model cenderung mengagregasikan pembahasan tersebut ke dalam topik yang lebih luas berdasarkan kedekatan semantik umum, bukan menangkapnya sebagai sub-isu yang spesifik. Misalnya, pembahasan yang mengaitkan aktor tertentu dengan tudingan ideologis atau afiliasi tertentu dapat terklaster bersama istilah bernada negatif lain yang lebih dominan dan sering muncul, sehingga detail isu tidak muncul secara eksplisit dalam representasi kata kunci topik.

Selain itu, terdapat kemungkinan bahwa isu-isu yang muncul secara organik tertutupi oleh volume percakapan lain yang bersifat repetitif, otomatis, atau diproduksi oleh akun non-organik. Meskipun penelitian ini telah mengupayakan reduksi *noise* melalui penyaringan anomali *burst* temporal, proses tersebut tidak sepenuhnya mengeliminasi seluruh distorsi wacana. Konsekuensinya, sebagian isu dengan intensitas tinggi namun leksikalnya tidak stabil

berpotensi teredam dalam struktur topik akhir. Temuan ini menunjukkan bahwa hasil pemodelan topik merepresentasikan pola diskursus dominan yang paling stabil secara semantik, bukan keseluruhan detail isu yang pernah muncul dalam percakapan publik.

Meskipun beberapa isu berpotensi teredam karena distorsi *burst*, hasil pemodelan tetap merefleksikan pola diskursus yang stabil. Hal ini menunjukkan bahwa pendekatan BERTopic, dengan kemampuannya menangkap struktur topik yang koheren, memberikan representasi yang lebih kaya dibanding analisis sentimen sederhana, sehingga relevan untuk penyusunan strategi komunikasi publik. Implikasinya, komunikasi kebijakan publik dapat disusun secara lebih sistematis, tidak hanya menekankan narasi makroekonomi berlandaskan polaritas emosi, tetapi juga merespons aspek lain seperti kebingungan publik dan kekhawatiran terhadap tata kelola. Dengan demikian, hasil pemodelan topik ini menyediakan dasar empiris yang relevan bagi pembuat kebijakan untuk memperbaiki strategi komunikasi, meningkatkan transparansi, dan memperkuat kepercayaan publik terhadap Danantara.

Perlu ditegaskan bahwa struktur diskursus yang dihasilkan dalam penelitian ini bersifat kontekstual terhadap wacana Danantara. Oleh karena itu, temuan dan implikasi yang dibahas tidak dimaksudkan untuk digeneralisasi secara langsung ke isu kebijakan lain tanpa evaluasi ulang terhadap karakteristik data dan konteks topiknya.

4.9 Kelebihan dan Kekurangan Penelitian

Penelitian ini memiliki sejumlah kelebihan metodologis yang relevan dalam konteks pemetaan wacana publik berbasis media sosial. Penggunaan BERTopic yang mengoptimasi parameter dalam UMAP dan HDBSCAN menggunakan Optuna, memungkinkan pembentukan struktur topik yang koheren, beragam, dan interpretatif. Namun, penetapan ruang pencarian *hyperparameter* dalam Optuna melibatkan keputusan heuristik peneliti yang didasarkan pada eksplorasi awal dan praktik umum pada dokumentasi resmi. Oleh karena itu, nilai numerik yang digunakan tidak dapat dianggap sebagai standar universal dan berpotensi menghasilkan struktur topik yang berbeda apabila diterapkan pada konteks data atau isu yang lain.

Penerapan pembersihan data berlapis, termasuk penghapusan duplikasi dan reduksi anomali temporal berbasis *burst*, berkontribusi dalam menurunkan dominasi percakapan yang bersifat repetitif dan tidak proporsional. Namun demikian, pendekatan *burst removal* yang digunakan dalam penelitian ini memiliki keterbatasan mendasar. Deteksi *burst* tidak dirancang untuk mengidentifikasi jaringan koordinasi antar akun, pola interaksi sosial, maupun hubungan

struktural antar pengguna, sehingga hasil penghapusan *burst* tidak dapat dipandang sebagai bukti keberadaan atau ketiadaan aktivitas *astroturfing* (Schoch dkk., 2022).

Lebih lanjut, mekanisme *burst removal* yang bertujuan mengurangi percakapan anomali atau non-organik secara inheren mengandung risiko terhapusnya percakapan organik yang sah. Risiko ini muncul karena identifikasi *burst* hanya didasarkan pada indikasi lonjakan frekuensi temporal yang tidak wajar, tanpa mempertimbangkan konteks sosial, intensi pengguna, maupun karakteristik akun secara individual. Keterbatasan data penelitian, khususnya tidak tersedianya informasi akun yang lebih rinci, menyebabkan peneliti tidak dapat melakukan analisis koordinasi antar akun atau validasi sosial terhadap *burst* yang terdeteksi. Selain itu, keterbatasan waktu penelitian juga membatasi kemungkinan pengumpulan data tambahan yang diperlukan untuk melakukan verifikasi lebih lanjut terhadap temuan *burst*.

Konsekuensi dari keterbatasan tersebut adalah bahwa potensi terbentuknya *echo chambers*, dominasi isu oleh kelompok tertentu, maupun dinamika penyebaran wacana berbasis jaringan sosial belum dapat dijelaskan dalam penelitian ini. Keterbatasan lain terletak pada ruang lingkup data yang hanya berasal dari satu platform media sosial, sehingga ketahanan dan konsistensi struktur topik pada konteks lintas platform tidak dapat diuji dan dibandingkan.

Dari sisi metodologis, perumusan fungsi objektif dalam proses *hyperparameter tuning* melibatkan penilaian manual dalam penentuan bobot metrik evaluasi, yang membuka ruang terjadinya bias peneliti, meskipun telah dilakukan evaluasi lintas konfigurasi model untuk meminimalkan pengaruh bias tersebut. Penamaan dan interpretasi topik sepenuhnya didasarkan pada kata kunci c-TF-IDF dan dokumen representatif yang dipilih oleh model, sehingga subjektivitas dalam pemaknaan topik tidak dapat dieliminasi sepenuhnya. Oleh karena itu, hasil penelitian ini bersifat deskriptif dan bertujuan memetakan struktur wacana, bukan menjelaskan mekanisme produksi, penyebaran, maupun dampak praktis wacana secara kausal.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini bertujuan untuk memetakan struktur wacana publik mengenai Danantara pada platform media sosial X melalui pendekatan pemodelan topik berbasis BERTopic yang dioptimasi melalui parameter-parameter UMAP dan HDBSCAN. Berdasarkan hasil analisis yang telah dipaparkan pada Bab IV, beberapa kesimpulan utama dapat dirumuskan sebagai berikut:

a. Analisis terhadap data percakapan publik menunjukkan bahwa karakteristik data dari media sosial X tercermin melalui keberadaan pola duplikasi, spam, serta anomali temporal yang signifikan pada data mentah, sehingga menegaskan perlunya tahapan pembersihan data yang sistematis sebelum pemodelan topik dilakukan.
b. Optimasi parameter UMAP dan HDBSCAN menggunakan Optuna menghasilkan konfigurasi pemodelan topik yang lebih stabil dan seimbang dibandingkan pendekatan tanpa optimasi. Proses perumusan fungsi objektif secara bertahap yang mempertimbangkan koherensi, keragaman, dan jumlah topik, menghasilkan model final dengan pemisahan antartopik yang jelas dan tetap koheren.
c. Pemodelan topik menghasilkan lima belas topik substantif yang merepresentasikan isu-isu tematik utama dalam wacana publik mengenai Danantara. Topik-topik tersebut mencakup optimisme ekonomi dan investasi, diskursus teknis terkait pengelolaan aset BUMN dan sektor strategis, kritik terhadap tata kelola dan transparansi, serta ekspresi kebingungan dan percakapan informal. Distribusi topik menunjukkan bahwa Danantara dipersepsikan sebagai entitas multidimensional yang berada pada irisan kebijakan ekonomi, politik, dan budaya komunikasi media sosial.

Secara keseluruhan, penelitian ini berhasil mencapai seluruh tujuan yang telah ditetapkan dengan menyajikan pemetaan deskriptif terhadap struktur wacana publik mengenai Danantara. Temuan yang dihasilkan dimaksudkan memberikan gambaran empiris mengenai bagaimana isu Danantara dibicarakan, dipersepsikan, dan didistribusikan dalam ruang diskursus publik digital selama periode pengamatan.

5.2 Saran

Berdasarkan hasil dan keterbatasan penelitian ini, beberapa saran dapat diajukan untuk pengembangan penelitian selanjutnya maupun pemanfaatan praktis temuan. Saran-saran tersebut dapat dilihat sebagai berikut:

a. Pertama, penelitian lanjutan disarankan untuk mengeksplorasi dinamika temporal topik secara lebih mendalam. Analisis pergeseran topik antarperiode pengamatan dapat memberikan pemahaman yang lebih rinci mengenai bagaimana wacana publik berkembang seiring dengan peristiwa kebijakan, pemberitaan media, atau dinamika politik tertentu yang berkaitan dengan Danantara.
b. Kedua, integrasi pemodelan topik dengan analisis jaringan interaksi akun di media sosial. Pendekatan ini berpotensi mengungkap pola pembentukan komunitas wacana, dominasi topik oleh kelompok tertentu, dan kemungkinan terbentuknya *echo chambers* atau aktivitas terkoordinasi. Hal ini akan mengakibatkan hasil pemodelan topik lebih organik dan terhindar dari isu-isu yang tercipta tidak alami.
c. Ketiga, validasi lintas platform dengan menerapkan *pipeline* analisis yang sama pada data dari media sosial lain, seperti Instagram, Facebook, atau YouTube, dapat dilakukan untuk menguji konsistensi dan ketahanan temuan. Perbandingan lintas platform akan membantu menilai sejauh mana struktur wacana publik mengenai Danantara.
d. Keempat, dari sisi praktis, hasil pemetaan topik ini dapat dimanfaatkan oleh pemangku kebijakan dan praktisi komunikasi publik sebagai dasar empiris dalam merancang strategi komunikasi yang lebih responsif. Keberadaan topik kebingungan publik dan kritik terhadap transparansi menunjukkan perlunya penyampaian informasi yang lebih jelas, terstruktur, dan mudah dipahami.

Dengan demikian, penelitian ini tidak hanya memberikan kontribusi teoretis pada kajian pemodelan topik dan analisis wacana digital berbahasa Indonesia, tetapi juga menawarkan nilai praktis sebagai alat bantu berbasis data untuk memahami dan mengelola persepsi publik terhadap kebijakan strategis negara seperti Danantara.

DAFTAR PUSTAKA

Albanese, N. C. (2022, September 19). *Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison | Towards Data Science*. towardsdatascience.com. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5

Andika, R., & Kusrini, K. (2025). Optimasi Hyperparameter Model LSTM dan Variannya Untuk Peramalan Pembelian Bahan Baku Karet Alam. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *10*(3), 2627–2639. https://doi.org/10.29100/jipi.v10i3.7567

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv*. http://arxiv.org/abs/2008.09470

Aryani, D., Lucia Kharisma, I., Sujjada, A., & Kamdan, K. (2024). Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, *9*(2), 171–180. https://doi.org/10.25139/inform.v9i2.8429

Badan Pengelola Investasi Daya Anagata Nusantara (BPI Danantara). (2025). *Danantara Indonesia | Beranda - Danantara Indonesia*. https://www.danantaraindonesia.co.id/id

Barreto, S., Moura, R., Carvalho, J., Paes, A., & Plastino, A. (2021). *Sentiment analysis in tweets: an assessment study from classical to modern text representation models*. http://arxiv.org/abs/2105.14373

BERTopic - Advanced Transformer-Based Topic Modeling. (t.t.). Diambil 26 Desember 2025, dari https://bertopic.com/

Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. *Political Analysis*, *31*(2), 235–256. https://doi.org/10.1017/pan.2022.10

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Dalam *Journal of Machine Learning Research* (Vol. 3).

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900

Daeni, F. I. M., Rachmarani, F. A., & Ilham, R. (2023). Pengaruh Buzzer Politik dalam Pemilu: Tantangan Terhadap Electoral Justice dalam Mempertahankan Prinsip Demokrasi. *Padjadjaran Law Review*, *11*(2), 191–211. https://doi.org/10.56895/plr.v11i2.1288

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Diana, K. D. R., & Khodra, M. L. (2023). IndoSBERT: Enhancing Indonesian Sentence Embeddings with Siamese Networks Fine-tuning. *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA59291.2023.10390469

Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15351-4

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Fahmi, I. (2025, Februari 25). *Analisis Percakapan Publik: Danantara*. DroneEmprit. https://pers.droneemprit.id/analisis-percakapan-publik-danantara/

García, S. A., Hung, E. S., & Mottareale, D. (2022). Astroturfing as a strategy for manipulating public opinion on Twitter during the pandemic in Spain. *Profesional de la Informacion*, *31*(3). https://doi.org/10.3145/epi.2022.may.10

Geeksforgeeks. (t.t.). *Understanding TF-IDF (Term Frequency-Inverse Document Frequency) - GeeksforGeeks*. Diambil 26 Desember 2025, dari https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

Goswami, A., Kumar, A., & Dhanya, P. (2024). Bursty Event Detection Model for Twitter. Dalam P. S. and S. V. V. and P. B. and M. A. R. and S. G. Devismes Stéphane and Mandal (Ed.), *Distributed Computing and Intelligent Technology* (hlm. 338–355). Springer Nature Switzerland. https://doi.org/doi.org/10.1007/978-3-031-50583-6_23

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. http://arxiv.org/abs/2203.05794

Guo, Y. T., Li, Q. Q., & Liang, C. S. (2024). The rise of nonnegative matrix factorization: Algorithms and applications. Dalam *Information Systems* (Vol. 123). Elsevier Ltd. https://doi.org/10.1016/j.is.2024.102379

Hermanto, M. L., Fathoni, Ardhillah, O., & Ibrahim, A. (2025). Analisis Sentimen Masyarakat terhadap Danantara di Platform X dengan Metode SVM. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(4), 6779–6785. https://doi.org/10.36040/jati.v9i4.14189

Ihsan, M., & Adnan, R. S. (2022). Media Sosial Twitter sebagai Ruang Publik Virtual (Studi Kasus Penolakan Omnibus Law). *Syntax Literate: Jurnal Ilmiah Indonesia*, *7*(3), 3255–3267. https://doi.org/10.36418/syntax-literate.v7i3.6612

Indriani, L. (2024). *Pemodelan Topik Saran Mahasiswa Pada Simak Unismuh Menggunakan BERTopic*. UNIVERSITAS MUHAMMADIYAH MAKASSAR.

Janssens, W., Bogaert, M., & Poel, D. Van den. (2025). *LLM-Assisted Topic Reduction for BERTopic on Social Media Data*. http://arxiv.org/abs/2509.19365

Kementerian Keuangan Republik Indonesia. (2025). *Presiden Prabowo Resmi Luncurkan Badan Pengelola Investasi Danantara*. https://www.kemenkeu.go.id/informasi-publik/publikasi/berita-utama/Presiden-Luncurkan-Danantara

Khadijah, U. N., & Cahyono, N. (2024). Analisis Topic Modelling Pariwisata Yogyakarta Menggunakan Latent Dirichlet Allocation (LDA). *The Indonesian Journal of Computer Science*, *13*(4). https://doi.org/10.33022/ijcs.v13i4.3816

Kwon, N., Yoo, Y., & Lee, B. (2024). Novel Curriculum Learning Strategy Using Class-Based TF-IDF for Enhancing Personality Detection in Text. *IEEE Access*, *12*, 87873–87882. https://doi.org/10.1109/ACCESS.2024.3417180

Li, Q., Chao, Y., Li, D., Lu, Y., & Zhang, C. (2023). *Event Detection from Social Media Stream: Methods, Datasets and Opportunities*.

Li, R., González-Pizarro, F., Xing, L., Murray, G., & Carenini, G. (2023). *Diversity-Aware Coherence Loss for Improving Neural Topic Models*. http://arxiv.org/abs/2305.16199

Listyawan, N. B. R., Setiawan, N. Y., & Saputra, M. C. (2017). Topic Modelling Pada Aktivitas Pengembangan Perangkat Lunak Menggunakan BERTopic. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *1*(1), 2548–2964. http://j-ptiik.ub.ac.id

Mahrunisa, T. El, Handoko, T. C., & Pribadi, R. (2025). *Danantara: Usulan Untuk Kemakmuran Generasi Masa Depan*. https://institutharkatnegeri.org/wp-content/uploads/2025/03/Policy-Brief-Danantara.pdf

McInnes, L. (t.t.-a). *Basic UMAP Parameters — umap 0.5.8 documentation*. Diambil 18 Januari 2026, dari https://umap-learn.readthedocs.io/en/latest/parameters.html

McInnes, L. (t.t.-b). *Parameter Selection for HDBSCAN* — hdbscan 0.8.1 documentation*. Diambil 18 Januari 2026, dari https://hdbscan.readthedocs.io/en/latest/parameter_selection.html

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426

Meaney, C., Stukel, T. A., Austin, P. C., Moineddin, R., Greiver, M., & Escobar, M. (2023). Quality indices for topic model selection and evaluation: a literature review and case study. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02216-1

Movahed, F. S., Berahmand, K., Sheikhpour, R., Li, Y., Pan, S., & Jalili, M. (2026). Nonnegative Matrix Factorization in Dimensionality Reduction: A Survey. *ACM Computing Surveys*, *58*(5), 1–41. https://doi.org/10.1145/3767726

Njurumana, A. F., Hariadi, F., & Uly, N. B. (2025). Analisis Sentimen Pengguna X Terkait Danantara Menggunakan Metode Support Vector Machine. *Riau Jurnal Teknik Informatika*, *4*(2). https://doi.org/10.30606/rjti.v4i2.3439

Oktafiandi, H. (2023). Implementasi LDA untuk Pengelompokan Topik Twitter Bertagar #Mypertamina. *JURNAL EKONOMI DAN TEKNIK INFORMATIKA*.

Pangastuti, S. S., Rohmatullayaly, E. N., & Najmi, N. (2025). Topic Modeling for User Feedback Dataset. *Communications in Mathematical Biology and Neuroscience*, *2025*. https://doi.org/10.28919/cmbn/8932

Perdana, R. S., Muliawati, T. H., & Alexandro, R. (2015). Bot Spammer Detection in Twitter Using Tweet Similarity and Time Interval Entropy. *Jurnal Ilmu Komputer dan Informasi*, *8*(1), 19. https://doi.org/10.21609/jiki.v8i1.280

Portuna, D. (2024). *Analisis Perbandingan Topic Modelling Latent Dirichlet Allocation dan Latent Semantic Analysis pada Berita DetikFinance*. UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA.

Priatna, A. V., Tasman, A. R. A., & Velentina, R. A. (2025). Mekanisme akuntabilitas sovereign wealth fund: perbandingan antara BPI Danantara dan GPFG Norwegia. *Jurnal Ekonomi dan Bisnis*, *14*(3).

Puspita, E., Shiddieq, D. F., & Roji, F. F. (2024). Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 481–489. https://doi.org/10.57152/malcom.v4i2.1204

Rahmanto, A. N. (2022). Media Sosial dan Persepsi Publik tentang Good Governance pada Pemerintah Daerah di Solo Raya. *Jurnal Ilmu Komunikasi*, *20*(1), 88. https://doi.org/10.31315/jik.v20i1.6433

raniaprillyap. (2020). *NLTK Tokenisasi - Stopwords - Stemming - Sastrawi*. Kaggle. https://www.kaggle.com/code/raniaprillyap/nltk-tokenisasi-stopwords-stemming-sastrawi/notebook#Stopwords

Riaz, A., Abdulkader, O., Ikram, M. J., & Jan, S. (2025). Exploring topic modelling: a comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity. *International Journal of Electrical and Computer Engineering (IJECE)*, *15*(2), 1933. https://doi.org/10.11591/ijece.v15i2.pp1933-1948

Schoch, D., Keller, F. B., Stier, S., & Yang, J. H. (2022). Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-08404-9

Singh, T., & Kumari, M. (2021). Burst: real-time events burst detection in social text stream. *Journal of Supercomputing*, *77*(10), 11228–11256. https://doi.org/10.1007/s11227-021-03717-4

Socs, B. (2018, November 29). *Latent Dirichlet Allocation (LDA) – School of Computer Science*. https://socs.binus.ac.id/2018/11/29/latent-dirichlet-allocation-lda/

Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, *73*. https://doi.org/10.1016/j.bspc.2021.103456

Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052405

Suriadi, H. (2025). Krisis Kepercayaan Masyarakat terhadap Lembaga Publik di Era Disinformasi Digital. *Journal of Social, Educational and Religious Studies*, *1*, 38–52. https://suriaacademicpress.com/ojs/index.php/sj-sers

Syahrial, S., & Afidh, R. P. F. (2024). Fine-Tuning Topic Modelling: A Coherence-Focused Analysis of Correlated Topic Models. *Infolitika Journal of Data Science*, *2*(2), 82–87. https://doi.org/10.60084/ijds.v2i2.236

Veantiny, I., Rumah Sukun, R., Aryanti, R., & Agustiani, S. (2023). *Analisis Sentimen Ulasan Publik Terhadap Pembentukan Danantara Sebagai Lembaga Investasi Menggunakan Algoritma Naïve Bayes*. *99*(99). https://doi.org/10.47065/tin.v99i99.9999

Wahyuni, W., Lestari, T. P., Apriliana, M., & Gumelta, R. (2025). Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments. *Journal of Applied Informatics and Computing (JAIC)*, *9*(4), 2548–6861. http://jurnal.polibatam.ac.id/index.php/JAIC

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *arXiv*. http://arxiv.org/abs/2402.05672

Weston, S. J., Shryock, I., Light, R., & Fisher, P. A. (2023). Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *6*(2). https://doi.org/10.1177/25152459231160105

Wijanto, M. C., Widiastuti, I., & Yong, H.-S. (2024). Topic Modeling for Scientific Articles: Exploring Optimal Hyperparameter Tuning in BERT. *International Journal on Advanced Science*, *14*(3).

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-Based Bag-of-Words Model for Text Classification. *IEEE Access*, *8*, 82641–82652. https://doi.org/10.1109/ACCESS.2020.2991074

Yunanto, F., Kasanova, K., Rudiyanto, M., & Haqiqi, M. N. (2024). Wacana Kritis Tentang Konstruksi Narasi Media dan Persepsi Mahasiswa Pada Pemilu Presiden 2024 di Universitas Madura. *JURNALISTRENDi : JURNAL LINGUISTIK, SASTRA, DAN PENDIDIKAN*, *9*(2), 310–317. https://doi.org/10.51673/jurnalistrendi.v9i2.2259

Zein, A., Wahyuni, S., Saragih, S., Andhini, R., & Khairunnisa. (2025). Persepsi Masyarakat terhadap Transparansi Anggaran dan Akuntabilitas Pemerintah Kota Medan. *OPTIMAL Jurnal Ekonomi dan Manajemen*, *5*(1), 344–356. https://doi.org/10.55606/optimal.v5i1.5891

LAMPIRAN

Lampiran A – *Stopwords* Khusus

yg	aja	ya
nih	sih	dong
deh	loh	kok
kan	nya	gua
gw	lu	loe
udah	ini	itu
tuh	kayak	dll
banget	makanya	gitu
si	oke	dan
untuk	gue	apa
tau	ga	ada
mau	di	gak
kalo	iki	ae
opo	sing	yo
wes	ra	nek
utk	bikin	pake
jd	dr	biar
tp	udh	sampe
emang	klo	krn
tdk	dgn	ini.
ya?	ini,	sdh
bgt	nggak	dah
dg	liat	jg