



INSTITUT PERTANIAN BOGOR

**EVALUASI KINERJA BERTOPIC DALAM PEMODELAN
TOPIK *MULTIDOMAIN* PADA TEKS BERITA
BERBAHASA INDONESIA**

TESIS

**WAWAN SAPUTRA
M0501241017**

**SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
PROGRAM STUDI STATISTIKA DAN SAINS DATA
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI TESIS, SUMBER INFORMASI, PENGGUNAAN AI, DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Evaluasi Kinerja BERTopic dalam Pemodelan Topik *Multidomain* pada Teks Berita Berbahasa Indonesia” merupakan salah satu syarat untuk memperoleh gelar magister Statistika dan Sains Data. Karya ini adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dalam penyusunan karya ini, saya menggunakan bantuan kecerdasan buatan ChatGPT dan Gemini untuk membantu penyusunan kode program, memberikan ide dan alternatif pendekatan analisis, mendukung penyempurnaan isi dan struktur penulisan, membantu perumusan nama topik berdasarkan kata/frasa dan dokumen representatif, serta membantu perancangan ringkasan visual. Setelah menggunakan alat/layanan tersebut, saya meninjau, memverifikasi, dan menyunting hasil sesuai kebutuhan serta bertanggung jawab penuh atas isi karya tugas akhir ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2026

Wawan Saputra
M0501241017

© Hak Cipta milik IPB, 2026
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

SOROTAN

WAWAN SAPUTRA. Evaluasi Kinerja BERTopic dalam Pemodelan Topik *Multidomain* pada Teks Berita Berbahasa Indonesia. Dibimbing oleh AGUS MOHAMAD SOLEH dan ASEP SAEFUDDIN.

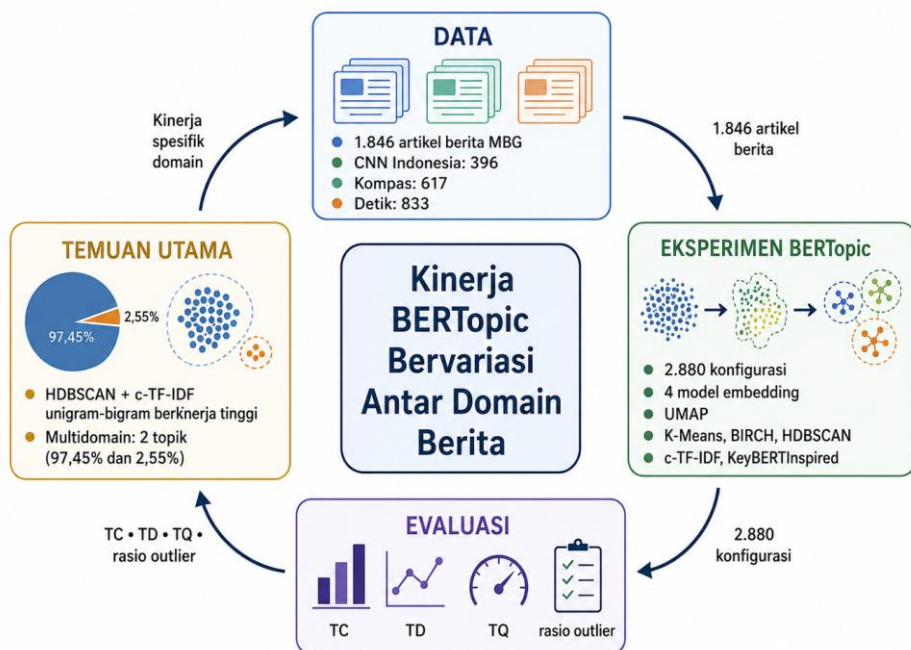
1. Penelitian ini mengevaluasi kinerja BERTopic pada berita daring *multidomain* berbahasa Indonesia.
2. Sebanyak 2.880 konfigurasi BERTopic dievaluasi menggunakan TC, TD, TQ, dan rasio outlier.
3. HDBSCAN dan c-TF-IDF unigram-bigram konsisten menghasilkan kinerja terbaik pada seluruh *domain*.
4. Kinerja model embedding berbeda antardomain sehingga konfigurasi terbaik bersifat spesifik *domain*.
5. Pemodelan *multidomain* menghasilkan struktur topik yang lebih ringkas dibandingkan *domain* tunggal.

HIGHLIGHTS

WAWAN SAPUTRA. *Performance Evaluation of BERTopic for Multidomain Topic Modeling of Indonesian News Texts. Supervised by AGUS MOHAMAD SOLEH and ASEP SAEFUDDIN.*

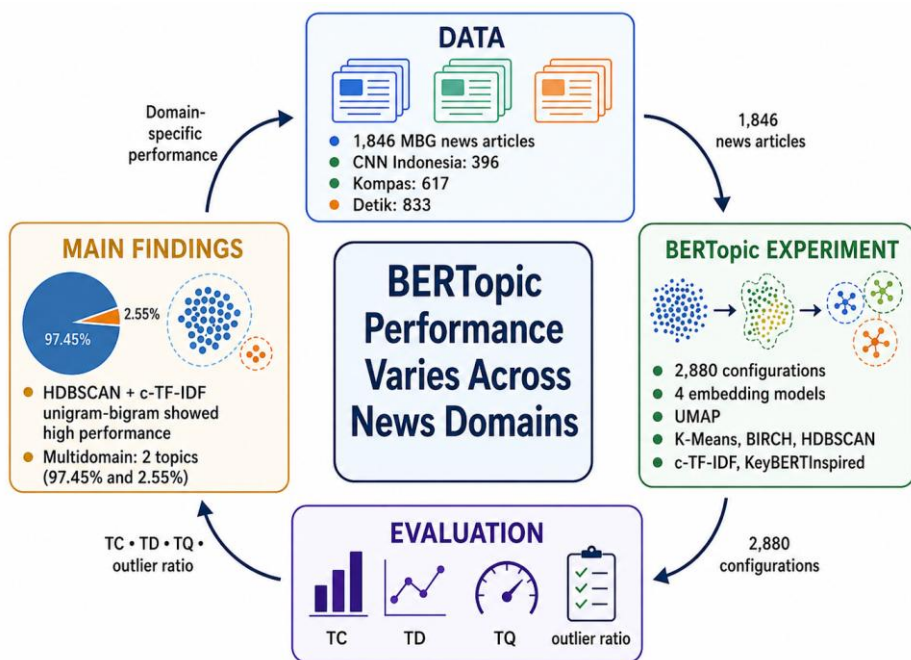
1. *This study evaluates BERTopic performance on multidomain Indonesian online news.*
2. *A total of 2,880 BERTopic configurations were evaluated using TC, TD, TQ, and outlier ratio.*
3. *HDBSCAN and c-TF-IDF unigram-bigram consistently achieved the best performance across domains.*
4. *Embedding model performance varied across domains, making the best configuration domain-specific.*
5. *Multidomain modeling produced a more compact topic structure than single-domain modeling.*

RINGKASAN VISUAL



Kinerja BERTopic Bervariasi Antardomain Berita

GRAPHICAL SUMMARY



BERTopic Performance Varies Across News Domains

RINGKASAN

WAWAN SAPUTRA. Evaluasi Kinerja BERTopic dalam Pemodelan Topik *Multidomain* pada Teks Berita Berbahasa Indonesia. Dibimbing oleh AGUS MOHAMAD SOLEH dan ASEP SAEFUDDIN.

Pemodelan topik berbasis Transformer semakin banyak digunakan untuk mengidentifikasi struktur tematik pada data teks berskala besar. Salah satu metode yang berkembang adalah BERTopic, yang mengintegrasikan representasi dokumen berbasis *embedding*, reduksi dimensi, *clustering*, dan representasi topik. Kinerja BERTopic dapat bervariasi berdasarkan kombinasi komponen dan parameter yang digunakan serta karakteristik korpus yang dianalisis. Kajian yang mengevaluasi BERTopic secara sistematis pada teks berita berbahasa Indonesia dari beberapa portal juga masih terbatas. Penelitian ini bertujuan mengevaluasi secara deskriptif kinerja BERTopic berdasarkan variasi model *embedding*, algoritma *clustering*, dan metode representasi topik, mengidentifikasi konfigurasi dengan TQ maksimum pada setiap *domain*, serta menginterpretasikan topik dan komposisi sumber berita pada pemodelan *Multidomain*.

Data penelitian terdiri atas 1.846 artikel berita mengenai Program Makan Bergizi Gratis (MBG) yang berasal dari CNN Indonesia sebanyak 396 artikel, Kompas sebanyak 617 artikel, dan Detik sebanyak 833 artikel. Data dianalisis dalam empat *domain*, yaitu CNN Indonesia, Kompas, Detik, dan *Multidomain* yang merupakan gabungan ketiga portal. Eksperimen membandingkan empat model *embedding*, yaitu IndoBERT, IndoSBERT, multilingual BERT, dan multilingual MPNet. *Embedding* direduksi menggunakan UMAP dengan *n_components* 5, 10, dan 15, kemudian dikelompokkan menggunakan K-Means, BIRCH, dan HDBSCAN dengan variasi parameter 5, 10, 15, 20, dan 30. Representasi topik dibandingkan menggunakan c-TF-IDF dan KeyBERTInspired dengan rentang unigram (1,1) dan unigram-bigram (1,2). Sebanyak 720 konfigurasi dievaluasi pada setiap *domain* sehingga diperoleh 2.880 konfigurasi. Evaluasi dilakukan menggunakan *Topic Coherence* (TC), *Topic Diversity* (TD), dan *Topic Quality* (TQ), sedangkan proporsi outlier digunakan sebagai informasi tambahan untuk menilai cakupan dokumen pada konfigurasi HDBSCAN.

Hasil menunjukkan bahwa kinerja komponen BERTopic bervariasi antardomain. Berdasarkan rata-rata TQ, IndoBERT menghasilkan nilai tertinggi pada CNN Indonesia, sedangkan IndoSBERT menghasilkan rata-rata tertinggi pada Kompas, Detik, dan *Multidomain*. HDBSCAN menghasilkan rata-rata dan nilai maksimum TQ tertinggi dibandingkan K-Means dan BIRCH pada seluruh *domain*. Demikian pula, c-TF-IDF dengan rentang unigram-bigram (1,2) menghasilkan rata-rata dan nilai maksimum TQ tertinggi pada seluruh *domain* dibandingkan metode representasi lainnya. Hasil tersebut menunjukkan bahwa model *embedding* bersifat lebih spesifik terhadap *domain*, sedangkan HDBSCAN dan c-TF-IDF unigram-bigram menunjukkan pola kinerja yang lebih konsisten dalam ruang konfigurasi yang dievaluasi.

Konfigurasi terbaik CNN Indonesia menggunakan IndoBERT, UMAP dengan *n_components* 15, HDBSCAN dengan *min_cluster_size* 30, dan c-TF-IDF unigram-bigram, dengan TQ sebesar 0,7489 dan proporsi outlier 58,59%. Konfigurasi terbaik Kompas menggunakan IndoBERT, UMAP-5, HDBSCAN



dengan *min_cluster_size* 15, dan c-TF-IDF unigram-bigram, dengan TQ tertinggi sebesar 0,7831 dan proporsi outlier 20,42%. Pada Detik, konfigurasi terbaik menggunakan IndoSBERT, UMAP-15, HDBSCAN dengan *min_cluster_size* 15, dan c-TF-IDF unigram-bigram, dengan TQ sebesar 0,7355 dan proporsi outlier 0,96%. Pada *multidomain*, beberapa konfigurasi multilingual MPNet mencapai TQ maksimum yang sama sebesar 0,7499. Salah satu konfigurasi tersebut menggunakan UMAP dengan *n_components* = 5, HDBSCAN dengan *min_cluster_size* = 15, dan c-TF-IDF unigram-bigram, serta menghasilkan TC sebesar 0,7499, TD sebesar 1,0000, dan tidak menghasilkan dokumen outlier. Perbedaan proporsi outlier tersebut menunjukkan bahwa TQ yang tinggi tidak selalu disertai cakupan dokumen yang tinggi sehingga kedua informasi perlu dipertimbangkan secara bersama dalam menafsirkan kinerja BERTopic.

Interpretasi terhadap konfigurasi yang dipilih menghasilkan tiga topik pada CNN Indonesia, yaitu keracunan makanan siswa, anggaran dan tata kelola MBG, serta wacana pendanaan zakat. Kompas menghasilkan dua topik, yaitu pelaksanaan MBG di sekolah dan wacana zakat untuk pendanaan MBG, sedangkan Detik menghasilkan dua topik berupa pelaksanaan dan isu keamanan pangan MBG serta mitra dapur dan pembayaran MBG. Pada *Multidomain* terbentuk dua topik, yaitu pelaksanaan umum program MBG dan susu, sapi perah, serta rantai pasok. Topik pelaksanaan umum mencakup 1.799 dari 1.846 dokumen dan memiliki komposisi portal yang relatif mengikuti komposisi awal korpus, sedangkan topik susu, sapi perah, dan rantai pasok terdiri atas 47 dokumen dengan proporsi dokumen Kompas yang relatif lebih besar. Hasil tersebut menunjukkan bahwa pemodelan *Multidomain* menghasilkan struktur topik yang lebih ringkas dan dapat menggambarkan tema bersama lintas portal, sedangkan pemodelan *domain* tunggal memberikan rincian tema yang tidak selalu terbentuk sebagai topik tersendiri setelah korpus digabungkan. Secara keseluruhan, hasil penelitian menunjukkan bahwa evaluasi BERTopic perlu mempertimbangkan kombinasi komponen pemodelan, kualitas topik, cakupan dokumen, dan karakteristik *domain*.

Kata kunci: *BERTopic*, *c-TF-IDF*, *embedding*, *HDBSCAN*, *Multidomain*

SUMMARY

WAWAN SAPUTRA. *Performance Evaluation of BERTopic for Multidomain Topic Modeling of Indonesian News Texts*. Supervised by AGUS MOHAMAD SOLEH and ASEP SAEFUDDIN.

Transformer-based topic modeling has increasingly been used to identify thematic structures in large-scale text corpora. One increasingly adopted approach is BERTopic, which integrates embedding-based document representation, dimensionality reduction, clustering, and topic representation. However, BERTopic performance may vary across different modeling components, parameter settings, and corpus characteristics. Systematic evaluations of BERTopic on Indonesian-language news texts collected from multiple news portals remain limited. This study aimed to descriptively evaluate BERTopic performance across different embedding models, clustering algorithms, and topic representation methods; identify the configurations achieving the maximum TQ in each domain; and interpret the resulting topics and news-source composition in Multidomain modeling.

The dataset consisted of 1,846 news articles concerning Indonesia's Free Nutritious Meal Program (Makan Bergizi Gratis, MBG), comprising 396 articles from CNN Indonesia, 617 from Kompas, and 833 from Detik. The data were analyzed in four domains: CNN Indonesia, Kompas, Detik, and a Multidomain corpus combining the three portals. The experiment compared four embedding models: IndoBERT, IndoSBERT, multilingual BERT, and multilingual MPNet. The embeddings were reduced using UMAP with $n_components$ of 5, 10, and 15 and subsequently clustered using K-Means, BIRCH, and HDBSCAN with parameter values of 5, 10, 15, 20, and 30. Topic representations were compared using c-TF-IDF and KeyBERTInspired with unigram (1,1) and unigram-bigram (1,2) ranges. A total of 720 configurations were evaluated for each domain, resulting in 2,880 configurations. Performance was assessed using Topic Coherence (TC), Topic Diversity (TD), and Topic Quality (TQ), while the outlier proportion was used as additional information to assess document coverage in HDBSCAN configurations.

The results showed that the performance of BERTopic components varied across domains. Based on mean TQ, IndoBERT achieved the highest value for CNN Indonesia, whereas IndoSBERT achieved the highest mean TQ for Kompas, Detik, and the Multidomain corpus. HDBSCAN produced the highest mean and maximum TQ values across all domains compared with K-Means and BIRCH. Similarly, c-TF-IDF with the unigram-bigram (1,2) range consistently achieved the highest mean and maximum TQ values across all domains compared with the other topic representation methods. These results indicate that embedding performance was more domain-specific, whereas HDBSCAN and c-TF-IDF with unigram-bigram representation exhibited more consistent performance patterns within the evaluated configuration space.

The best CNN Indonesia configuration used IndoBERT, UMAP with $n_components = 15$, HDBSCAN with $min_cluster_size = 30$, and c-TF-IDF unigram-bigram, achieving a TQ of 0.7489 with an outlier proportion of 58.59%. The best Kompas configuration used IndoBERT, UMAP-5, HDBSCAN with $min_cluster_size = 15$, and c-TF-IDF unigram-bigram, achieving the highest



overall TQ of 0.7831 with an outlier proportion of 20.42%. For Detik, the best configuration used IndoSBERT, UMAP-15, HDBSCAN with `min_cluster_size = 15`, and c-TF-IDF unigram-bigram, producing a TQ of 0.7355 and an outlier proportion of 0.96%. In the multidomain corpus, multiple multilingual MPNet configurations achieved the same maximum TQ of 0.7499. One of these configurations used UMAP with `n_components = 5`, HDBSCAN with `min_cluster_size = 15`, and c-TF-IDF unigram-bigram, producing a TC of 0.7499, a TD of 1.0000, and no outlier documents. The differences in outlier proportions demonstrate that a high TQ does not necessarily correspond to high document coverage. Therefore, both aspects should be considered when interpreting BERTopic performance.

Interpretation of the selected configurations identified three topics in CNN Indonesia: student food poisoning, MBG budgeting and governance, and the discourse on zakat-based funding. Kompas produced two topics: MBG implementation in schools and the discourse on using zakat to fund MBG. Detik also produced two topics: MBG implementation and food-safety issues, and kitchen partners and MBG payments. The Multidomain corpus produced two topics: general implementation of the MBG program and milk, dairy cattle, and the supply chain. The general implementation topic contained 1,799 of the 1,846 documents and had a source composition relatively similar to that of the overall corpus, whereas the milk, dairy cattle, and supply-chain topic contained 47 documents and had a relatively higher proportion of documents from Kompas. These findings indicate that Multidomain modeling produces a more compact topic structure, whereas single-domain modeling provides more detailed themes that do not necessarily remain distinct after the corpora are combined. Overall, the findings indicate that BERTopic evaluation should consider combinations of modeling components, topic quality, document coverage, and domain characteristics.

Keywords: BERTopic, c-TF-IDF, embedding, HDBSCAN, Multidomain



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Tesis:

1 Cici Suhaeni, S.Si., M.Si., Ph.D.



Judul Tesis : Evaluasi Kinerja BERTopic dalam Pemodelan Topik *Multidomain* pada Teks Berita Berbahasa Indonesia

Nama : Wawan Saputra

NIM : M0501241017

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Dr. Agus Mohamad Soleh, S.Si., M.T.

Pembimbing 2:

Prof. Dr. Ir. Asep Saefuddin, M.Sc.

Diketahui oleh

Ketua Program Studi :

Dr. Agus Mohamad Soleh, S.Si., M.T.

NIP 19750315 199903 1 004

Dekan Sekolah Sains Data, Matematika
dan Informatika:

Prof. Dr. Ir. Agus Buono, M.Si., M.Kom

NIP 19660702 199302 1 001

Tanggal Ujian: 24 Agustus 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Penelitian ini dilaksanakan sejak bulan Desember 2025 sampai bulan Mei 2026 dengan judul “Evaluasi Kinerja BERTopic dalam Pemodelan Topik *Multidomain* pada Teks Berita Berbahasa Indonesia”.

Terima kasih penulis ucapkan kepada para pembimbing, Dr. Agus Mohamad Soleh, S.Si., M.T. dan Prof. Dr. Ir. Asep Saefuddin, M.Sc. atas bimbingan dan nasihatnya sejak awal penelitian ini. Ucapan terima kasih juga penulis sampaikan kepada moderator seminar dan penguji luar komisi pembimbing. Penghargaan penulis sampaikan kepada seluruh dosen dan tenaga administrasi Program Studi Statistika dan Sains Data IPB University yang telah memberikan ilmu, dukungan, dan kemudahan selama proses pembelajaran pada jenjang magister.

Penulis mengucapkan terima kasih kepada Bapak dan Ibu atas perjuangan dalam membiayai pendidikan penulis hingga jenjang magister serta atas dukungan, doa, dan cinta yang tidak terhingga. Penulis juga mengucapkan terima kasih kepada teman dekat penulis yang telah banyak membantu selama perkuliahan, teman-teman Statistika dan Sains Data Angkatan 2024 yang selalu mendukung dan memberi semangat selama penyusunan tesis.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2026

Wawan Saputra



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



DAFTAR ISI

DAFTAR ISI.....	xi
DAFTAR TABEL.....	xii
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	xii
I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan.....	3
1.4 Manfaat.....	4
II TINJAUAN PUSTAKA.....	5
2.1 Pemodelan Topik.....	5
2.2 BERTopic.....	5
2.3 <i>Transformer</i> dan Representasi Teks.....	6
2.4 Model Embedding.....	7
2.5 <i>Uniform Manifold Approximation and Projection</i>	9
2.6 Algoritma <i>Clustering</i>	10
2.7 Representasi Topik.....	12
2.8 Evaluasi Kualitas Topik.....	13
III METODE.....	15
3.1 Data Penelitian.....	15
3.2 Rancangan Eksperimen BERTopic.....	15
3.3 Prosedur Analisis Data.....	17
3.4 Evaluasi dan Pemilihan Konfigurasi Terbaik.....	18
IV HASIL DAN PEMBAHASAN.....	20
4.1 Eksplorasi Data.....	20
4.2 Perbandingan <i>Embedding</i>	21
4.3 Perbandingan Algoritma <i>Clustering</i>	23
4.4 Perbandingan Representasi Topik.....	25
4.5 Konfigurasi Terbaik BERTopic.....	27
4.6 Interpretasi Topik.....	29
4.7 Analisis <i>Multidomain</i>	30
V SIMPULAN DAN SARAN.....	32
5.1 Simpulan.....	32
5.2 Saran.....	32
DAFTAR PUSTAKA.....	33
LAMPIRAN.....	36
RIWAYAT HIDUP.....	47



DAFTAR TABEL

1	Rancangan konfigurasi eksperimen BERTopic	16
2	Contoh praproses data	20
3	Statistik <i>topic quality</i> berdasarkan <i>domain</i> dan model <i>embedding</i>	22
4	Statistik <i>topic quality</i> berdasarkan <i>domain</i> dan algoritma <i>clustering</i>	24
5	Statistik <i>topic quality</i> berdasarkan <i>domain</i> dan representasi topik	26
6	Konfigurasi terbaik berdasarkan TQ	27
7	Topik utama hasil BERTopic pada konfigurasi terbaik	29
8	Komposisi <i>domain</i> pada data <i>Multidomain</i>	30

DAFTAR GAMBAR

1	Arsitektur Transformer (Vaswani <i>et al.</i> 2017)	6
2	Diagram alir penelitian	19
3	Visualisasi <i>word cloud domain</i> : (a) CNN Indonesia, (b) Kompas, (c) Detik, (d) <i>Multidomain</i>	21
4	Distribusi <i>topic quality</i> berdasarkan model <i>embedding</i> dan <i>domain</i>	23
5	Distribusi <i>topic quality</i> berdasarkan algoritma <i>clustering</i> dan <i>domain</i>	25
6	Distribusi <i>topic quality</i> berdasarkan representasi topik dan <i>domain</i>	27

DAFTAR LAMPIRAN

1	Contoh data berita penelitian	37
2	Diagram batang kata dominan pada setiap <i>domain</i>	38
3	Hasil konfigurasi CNN Indonesia	42
4	Hasil konfigurasi Kompas	42
5	Hasil konfigurasi Detik	43
6	Hasil konfigurasi <i>Multidomain</i>	43
7	Kata/Frasa representatif topik terbaik	44