PEMODELAN TOPIK BERITA *ONLINE* MENGGUNAKAN BERTOPIC DENGAN *SENTENCE EMBEDDING* INDOSBERT DAN MULTILINGUAL MPNET

CINDY INDRIYANI

PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

IPB University
Bogor Indonesia
@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.
IPB University
Perpustakaan IPB University

@Hak cipta milik IPB University

IPB University

IPB University
Bogor Indonesia

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.

Perpustakaan IPB University

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
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.
Perpustakaan IPB University

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul "Pemodelan Topik Berita *Online* Menggunakan BERTopic dengan *Sentence Embedding* IndoSBERT dan Multilingual MPNet" 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 skripsi ini.

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

Bogor, Juni 2026

Cindy Indriyani
G1401221095

IPB University
Bogor Indonesia

ABSTRAK

CINDY INDRIYANI. Pemodelan Topik Berita *Online* Menggunakan BERTopic dengan *Sentence Embedding* IndoSBERT dan Multilingual MPNet. Dibimbing oleh RAHMA ANISA dan LAILY NISSA ATUL MUALIFAH.

Perkembangan media digital menghasilkan volume berita *online* besar dan tidak terstruktur, sehingga memerlukan pemetaan isu otomatis guna menangkap arah kebijakan publik. Metode yang relevan adalah BERTopic yang mengandalkan model *embedding* sebagai komponen penentu representasi makna dokumen. Oleh karena itu, penelitian ini membandingkan model *sentence embedding* IndoSBERT yang mampu menangkap karakteristik kosakata lokal dengan multilingual MPNet sebagai model lintas bahasa. Tujuan penelitian ini mencakup analisis karakteristik *embedding* IndoSBERT dan multilingual MPNet, evaluasi pengaruh optimasi *hyperparameter* berbasis Optuna pada model BERTopic, serta identifikasi topik dominan dan dinamika temporal pemberitaan. Penelitian ini menggunakan detikcom sebagai sumber data representatif dinamika isu di masyarakat. Data yang dianalisis mencakup 169.446 berita periode Januari 2023 hingga Desember 2025. *Dataset* diproses menggunakan BERTopic dengan integrasi UMAP, K-Means++, dan *hierarchical clustering* sebagai alat interpretasi. Hasil analisis menunjukkan bahwa IndoSBERT lebih baik dalam menangkap konteks kalimat tunggal berbahasa Indonesia secara mendalam karena efisiensi proses tokenisasi serta keunggulan arsitektur modelnya. Melalui optimasi Optuna, BERTopic-IndoSBERT sebagai model terbaik mencapai skor *topic quality* (TQ) sebesar 0,5548 dengan 38 topik. Melalui pendekatan hierarki, topik tersebut terorganisasi ke dalam sembilan klaster dengan dominasi klaster Kebijakan Publik dan Tata Kelola Pemerintahan (24,15%). Analisis ini menangkap pergeseran fokus pemberitaan dari kompetisi politik menuju implementasi kebijakan pemerintah.

Kata kunci: BERTopic, detikcom, *embedding*, IndoSBERT, MPNet

@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.
IPB University
Perpustakaan IPB University

IPB University
Bogor Indonesia
@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.
IPB University
Perpustakaan IPB University

ABSTRACT

CINDY INDRIYANI. Topic Modeling of Online News Using BERTopic with IndoSBERT and Multilingual MPNet Sentence Embeddings. Supervised by RAHMA ANISA and LAILY NISSA ATUL MUALIFAH.

The development of digital media has resulted in a large and unstructured volume of online news, necessitating automated issue mapping to capture public policy directions. A relevant method is BERTopic, which relies on an embedding model as a determining component in document meaning representation. Therefore, this study compared the IndoSBERT sentence embedding model, which is capable of capturing local vocabulary characteristics, with the multilingual MPNet as a cross-language model. The objectives of this study included analyzing the characteristics of IndoSBERT and multilingual MPNet embeddings, evaluating the effect of Optuna-based hyperparameter optimization on the BERTopic model, and identifying dominant topics and the temporal dynamics of news coverage. This study used detikcom as a representative data source of issue dynamics in society. The analyzed data included 169.446 news items from January 2023 to December 2025. The dataset was processed using BERTopic with the integration of UMAP, K-Means++, and hierarchical clustering as interpretation tools. The analysis results showed that IndoSBERT was better at capturing the context of single Indonesian sentences in depth due to the efficiency of the tokenization process and the superiority of its model architecture. Through Optuna optimization, BERTopic-IndoSBERT, as the best model, achieved a topic quality (TQ) score of 0,5548 with 38 topics. Using a hierarchical approach, these topics were organized into nine clusters, dominated by the Public Policy and Governance cluster (24,15%). This analysis captured the shift in news coverage focus from political competition to government policy implementation.

Keywords: BERTopic, detikcom, embedding, IndoSBERT, MPNet

IPB University
Bogor Indonesia
@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.

© Hak Cipta milik IPB, tahun 2026[1]
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.

IPB University
Perpustakaan IPB University

PEMODELAN TOPIK BERITA *ONLINE* MENGGUNAKAN BERTOPIC DENGAN *SENTENCE EMBEDDING* INDOSBERT DAN MULTILINGUAL MPNET

CINDY INDRIYANI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Statistika dan Sains Data

PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

IPB University
Bogor Indonesia
@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.
IPB University
Perpustakaan IPB University

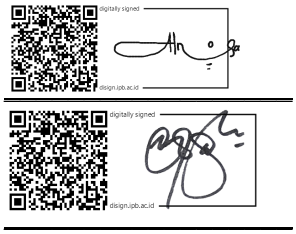
IPB University
Bogor Indonesia
@Hak cipta milik IPB University
Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, 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.
IPB University
Perpustakaan IPB University

Tim Penguji pada Ujian Skripsi:
1. Ir. Aam Alamudi, M.Si.

Judul Skripsi : Pemodelan Topik Berita *Online* Menggunakan BERTopic dengan *Sentence Embedding* IndoSBERT dan Multilingual MPNet
Nama : Cindy Indriyani
NIM : G1401221095

Disetujui oleh

Pembimbing 1:
Rahma Anisa, S.Stat., M.Si., M.Act.Sc.



Pembimbing 2:
Laily Nissa Atul Mualifah, S.Si., M.Si.

Diketahui oleh

Ketua Program Studi:
Dr. Bagus Sartono, S.Si., M.Si.
NIP 197804112005011002

Tanggal Ujian:
25 Mei 2026

Tanggal Lulus:

IPB University
Bogor Indonesia
@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.
IPB University
Perpustakaan IPB University

PRAKATA

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat dan karunia-Nya, sehingga karya ilmiah ini dapat diselesaikan dengan baik. Penelitian ini disusun dengan judul "Pemodelan Topik Berita *Online* Menggunakan BERTopic dengan *Sentence Embedding* IndoSBERT dan Multilingual MPNet". Penulisan skripsi ini tidak lepas dari dukungan berbagai pihak yang telah membantu penulis sejak masa studi hingga selesainya karya ilmiah ini. Oleh karena itu, penulis ingin menyampaikan terima kasih yang tulus kepada:

1. Kedua orang tua penulis, Bapak Juharno dan Ibu Rosinah, serta Emay Mayang Sari dan seluruh keluarga besar yang senantiasa memberikan doa, dukungan moril, serta kasih sayang yang menjadi kekuatan terbesar penulis.
2. Ibu Rahma Anisa, S.Stat., M.Si., M.Act.Sc. dan Ibu Laily Nissa Atul Mualifah, S.Si., M.Si. selaku komisi pembimbing yang telah meluangkan waktunya untuk memberikan bimbingan, arahan, dan ilmu selama penyusunan karya ilmiah ini.
3. Bapak Ir. Aam Alamudi, M.Si. selaku dosen penguji yang telah memberikan kritik serta saran membangun demi kesempurnaan penelitian ini.
4. Seluruh Bapak dan Ibu Dosen serta tenaga kependidikan Program Studi Statistika dan Sains Data IPB University yang telah membekali penulis dengan ilmu pengetahuan dan bantuan administratif selama masa perkuliahan.
5. Teman-teman seperjuangan Statistika Angkatan 59, khususnya Deswita Nur Alpharofi, Adinda Pratiwi, Yulianti Nurdzanah, dan Siti Arbaynah yang selalu menemani, memberikan semangat, serta menjadi rekan diskusi yang luar biasa selama masa studi.
6. Zulfa Alkomariah, Dea Azzahra, Syifa Nailah Khansa, dan Vera Nur Fadiya sebagai sahabat yang telah memberikan dukungan dan semangat bagi penulis sejak masa sekolah dasar hingga saat ini.
7. Rekan-rekan dari organisasi, terutama Birena Sinar 59 yang telah memberikan pengalaman berharga dan semangat dalam melewati masa-masa kuliah yang menantang.
8. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah membantu baik secara langsung maupun tidak langsung dalam penyelesaian skripsi ini.

Penulis menyadari bahwa karya ilmiah ini masih jauh dari sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun dari pembaca. Semoga penelitian ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan, khususnya di bidang statistika dan sains data.

Bogor, Juni 2026

Cindy Indriyani

IPB University
Bogor Indonesia
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.
© Hak cipta milik IPB University
IPB University
Perpustakaan IPB University

IPB University
Bogor Indonesia
@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.
IPB University
Perpustakaan IPB University

DAFTAR ISI

DAFTAR TABEL ix
DAFTAR GAMBAR ix
DAFTAR LAMPIRAN x
I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan 3
II TINJAUAN PUSTAKA 4
2.1 BERTopic 4
2.2 *Sentence Embedding* 4
2.2.1 *Sentence* BERT 4
2.2.2 Model IndoSBERT 5
2.2.3 Model Multilingual Berbasis MPNet 6
2.3 Normalisasi L2 6
2.4 *Uniform Manifold Approximation and Projection* (UMAP) 6
2.5 K-Means++ 7
2.6 *Class-Based Term Frequency Invers Document Frequency* (c-TF-IDF) 8
2.7 Optuna 8
2.8 Skor Koherensi Topik (*Topic Coherence*) 9
2.9 Skor Keragaman Topik (*Topic Diversity*) 9
2.10 *Hierarchical Clustering* 10
2.11 *Dunn Index* 10
III METODE 11
3.1 Data 11
3.2 Prosedur Analisis Data 11
IV HASIL DAN PEMBAHASAN 15
4.1 Pra-Pemrosesan Data 15
4.2 Eksplorasi Data 15
4.3 Hasil *Embedding* IndoSBERT 16
4.4 Hasil *Embedding* Multilingual MPNet 17
4.5 Hasil Pemodelan Topik BERTopic 18
4.6 Interpretasi Model Terbaik 22
4.6.1 *Hierarchical Clustering* 22
4.6.2 Dinamika Klaster Topik secara Temporal 24
4.6.3 Evolusi *Term* Setiap Klaster Topik 26
V SIMPULAN DAN SARAN 32
5.1 Simpulan 32
5.2 Saran 32
DAFTAR PUSTAKA 33
LAMPIRAN 37
RIWAYAT HIDUP 55

IPB University
Bogor Indonesia

DAFTAR TABEL

1	Rincian peubah data berita *online* detikcom hasil *scraping*	11
2	Ruang pencarian *hyperparameter* Optuna	13
3	Contoh teks berita sebelum dan sesudah pra-pemrosesan data	15
4	Contoh vektor *contextual embedding* pada IndoSBERT	16
5	Contoh vektor *contextual embedding* pada multilingual MPNet	17
6	Perbandingan hasil evaluasi model BERTopic	20
7	Sebaran dokumen hasil pemodelan BERTopic-IndoSBERT (Optuna)	22
8	Sebaran dokumen pada klaster topik hasil *hierarchical clustering*	23

DAFTAR GAMBAR

1	Prosedur analisis data	14
2	Distribusi berita detikcom periode 2023 hingga 2025	16
3	Perubahan skor kualitas topik pada model BERTopic dengan *embedding* IndoSBERT (a) dan multilingual MPNet (b) pada skema *baseline*	18
4	Perubahan skor kualitas topik pada model BERTopic dengan *embedding* IndoSBERT (a) dan multilingual MPNet (b) pada skema Optuna	20
5	Dinamika klaster topik pada model BERTopic-IndoSBERT (Optuna)	25
6	Evolusi *term* pada klaster topik Dinamika Politik Elektoral Nasional di tahun 2023 (a), 2024 (b), dan 2025 (c)	26
7	Evolusi *term* pada klaster topik Kebijakan Publik dan Tata Kelola Pemerintahan di tahun 2023 (a), 2024 (b), dan 2025 (c)	27
8	Evolusi *term* pada klaster topik Fenomena Kriminalitas dan Kejahatan di tahun 2023 (a), 2024 (b), dan 2025 (c)	27
9	Evolusi *term* pada klaster topik Penegakan Hukum dan Proses Demokrasi di tahun 2023 (a), 2024 (b), dan 2025 (c)	28
10	Evolusi *term* pada klaster topik Pengembangan Infrastruktur dan Mobilitas Kota di tahun 2023 (a), 2024 (b), dan 2025 (c)	29
11	Evolusi *term* pada klaster topik Manajemen Bencana dan Risiko Lingkungan di tahun 2023 (a), 2024 (b), dan 2025 (c)	29
12	Evolusi *term* pada klaster topik Dinamika Geopolitik Global di tahun 2023 (a), 2024 (b), dan 2025 (c)	30
13	Evolusi *term* pada klaster topik Informasi Meteorologi dan Geofisika di tahun 2023 (a), 2024 (b), dan 2025 (c)	30
14	Evolusi *term* pada klaster topik Aktivitas Sosial Keagamaan di tahun 2023 (a), 2024 (b), dan 2025 (c)	31

@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.
IPB University
Perpustakaan IPB University

DAFTAR LAMPIRAN

1	Hasil pemodelan topik BERTopic-IndoSBERT pada skema *baseline*	38
2	Hasil pemodelan topik BERTopic-multilingual MPNet pada skema *baseline*	42
3	Hasil pemodelan topik BERTopic-IndoSBERT pada skema Optuna	47
4	Hasil pemodelan topik BERTopic-multilingual MPNet pada skema Optuna	51
5	*Dendrogram* pengelompokan topik pada model BERTopic-IndoSBERT (Optuna)	53
6	Perbandingan nilai *dunn index* pada berbagai jumlah klaster (*k*)	54

IPB University
Bogor Indonesia
@Hak cipta milik IPB University
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.
Perpustakaan IPB University

@Hak cipta milik IPB University

IPB University

IPB University
Bogor Indonesia

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.

Perpustakaan IPB University