

METODE LDA DAN BERTOPIC DENGAN OPTIMASI HYPERPARAMETER HDBSCAN PADA WACANA PUBLIK IPB UNIVERSITY DI X

SABRINA ADNIN KAMILA



**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
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 DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Metode LDA dan BERTopic dengan Optimasi *Hyperparameter* HDBSCAN pada Wacana Publik IPB University di X” 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 proses penyusunan naskah tesis ini, penulis memanfaatkan kecerdasan artifisial (*Artificial Intelligence/AI*) generatif secara terbatas untuk membantu penyuntingan bahasa, perbaikan tata bahasa, dan penyempurnaan redaksi. Seluruh ide penelitian, metodologi, analisis data, hasil, pembahasan, dan simpulan sepenuhnya merupakan hasil pemikiran dan tanggung jawab penulis.

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

Bogor, Agustus 2026

Sabrina Adnin Kamila
M0501241042



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
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.

RINGKASAN

SABRINA ADNIN KAMILA. Metode LDA dan BERTopic dengan Optimasi *Hyperparameter* HDBSCAN pada Wacana Publik IPB University di X. Dibimbing oleh UTAMI DYAH SYAFITRI dan CICI SUHAENI.

Media sosial X menjadi ruang penting bagi wacana publik mengenai reputasi perguruan tinggi termasuk IPB University. *Latent Dirichlet Allocation* (LDA) merupakan metode *topic modeling* klasik yang banyak digunakan namun performanya menurun pada teks pendek dan informal seperti cuitan karena pola kemunculan bersama kata menjadi jarang. BERTopic hadir sebagai metode *topic modeling* berbasis *transformer* yang mengatasi keterbatasan tersebut melalui representasi semantik kontekstual. Meskipun demikian kinerja BERTopic sangat bergantung pada konfigurasi *hyperparameter* komponen klusterisasinya yaitu *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Konfigurasi *default* HDBSCAN dapat menghasilkan proporsi *outlier* yang tinggi dan koherensi topik yang rendah. Interaksi antara strategi *preprocessing* teks dan perilaku klusterisasi HDBSCAN juga belum banyak dikaji secara sistematis terutama untuk teks media sosial berbahasa Indonesia yang memiliki karakteristik linguistik khas.

Penelitian ini bertujuan mengevaluasi performa BERTopic yang telah dioptimalkan dibandingkan dengan BERTopic berkonfigurasi *default* dan LDA. Penelitian ini juga menganalisis pengaruh interaksi antara strategi *preprocessing* dan konfigurasi *hyperparameter* HDBSCAN terhadap kualitas dan struktur topik serta mengidentifikasi tema dominan dalam wacana publik mengenai reputasi IPB University di media sosial X. Data penelitian berupa 21.298 cuitan berbahasa Indonesia yang dikumpulkan pada periode Agustus 2023 hingga Agustus 2025. Data tersebut diolah melalui tiga skenario *preprocessing* yang berbeda dalam penanganan token numerik karakter tunggal dan penggabungan kata majemuk. Setiap skenario dimodelkan menggunakan LDA dan BERTopic dengan optimasi BERTopic dilakukan melalui *grid search* atas 126 kombinasi *hyperparameter* HDBSCAN pada setiap skenario. Pemilihan model terbaik menggabungkan evaluasi kuantitatif berupa *topic coherence*, *topic uniqueness*, *document coverage* dan *factuality* dengan pemeriksaan kualitatif terhadap *intertopic distance map*.

Hasil penelitian menunjukkan bahwa pengaruh *hyperparameter* HDBSCAN terhadap kualitas topik tidak seragam pada seluruh skenario *preprocessing*. Di antara keempat *hyperparameter* yang diuji *cluster_selection_method* terbukti paling berpengaruh terutama dalam menekan rasio *outlier*. Metode *eom* secara konsisten menghasilkan rasio *outlier* yang lebih rendah dibandingkan metode *leaf* dan perbedaan ini terbukti signifikan secara statistik berdasarkan uji Wilcoxon *signed-rank* berpasangan pada seluruh skenario. Sementara itu *cluster_selection_epsilon* tidak menunjukkan pengaruh yang berarti terhadap *topic coherence*. Konfigurasi *hyperparameter* optimal juga berbeda pada setiap skenario *preprocessing* bahkan metode *cluster_selection_method* terbaik berbalik pada skenario yang menerapkan penggabungan kata akibat pergeseran ruang *embedding*.

Optimasi *hyperparameter* HDBSCAN melalui *grid search* meningkatkan *topic coherence* BERTopic sebesar 0,148 hingga 0,251 dibandingkan konfigurasi *default* disertai berkurangnya jumlah topik sehingga lebih mudah diinterpretasikan.

BERTopic teroptimasi juga secara konsisten menghasilkan *topic coherence* yang lebih tinggi daripada LDA pada seluruh skenario sedangkan LDA justru unggul pada *topic uniqueness* dan *document coverage*. Kedua model sama-sama menghasilkan nilai *factuality* sebesar 1,000 pada seluruh skenario. Secara kualitatif BERTopic membentuk topik yang lebih terpisah dengan batas tema yang lebih jelas sementara LDA mampu mencakup proporsi dokumen yang lebih besar.

Penelitian ini berhasil memetakan enam tema dominan dalam wacana publik mengenai IPB University, yaitu penerimaan mahasiswa baru, ekspresi emosional calon mahasiswa dan mahasiswa baru, kehidupan akademik dan identitas kampus, biaya pendidikan, fasilitas dan infrastruktur kampus, serta peran IPB University dalam sektor pertanian dan ketahanan pangan. Penerimaan mahasiswa baru merupakan tema paling dominan yang konsisten muncul pada seluruh skenario sedangkan beberapa tema lain menunjukkan sensitivitas terhadap strategi *preprocessing* yang diterapkan. Hasil penelitian ini memberikan kerangka optimasi BERTopic yang dapat direplikasi untuk analisis wacana media sosial berbahasa Indonesia sekaligus memberikan gambaran bagi IPB University dalam menyusun strategi komunikasi dan pengelolaan reputasi yang lebih berbasis data.

Kata kunci: BERTopic, HDBSCAN, *hyperparameter*, LDA, *topic modeling*

SUMMARY

SABRINA ADNIN KAMILA. LDA and BERTopic Methods with HDBSCAN Hyperparameter Optimization on Public Discourse about IPB University on X. Supervised by UTAMI DYAH SYAFITRI and CICI SUHAENI.

Social media platform X has become an important space for public discourse on university reputation including that of IPB University. Latent Dirichlet Allocation (LDA) is a widely used classical topic modeling method yet its performance declines on short and informal texts such as tweets because word co-occurrence patterns become sparse. BERTopic emerged as a transformer-based topic modeling method that addresses this limitation through contextual semantic representation. Its performance nevertheless depends heavily on the hyperparameter configuration of its clustering component namely Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). A default HDBSCAN configuration can produce a high proportion of outliers and low topic coherence. The interaction between text preprocessing strategies and HDBSCAN clustering behavior has also not been widely studied systematically particularly for Indonesian-language social media text with distinctive linguistic characteristics.

This study aims to evaluate the performance of optimized BERTopic compared to default-configuration BERTopic and LDA. It also analyzes the interaction between preprocessing strategies and HDBSCAN hyperparameter configurations on topic quality and structure and identifies dominant themes in public discourse about IPB University's reputation on X. The dataset consisted of 21,298 Indonesian-language tweets collected from August 2023 to August 2025. The data were processed through three preprocessing scenarios that differed in their treatment of numeric tokens single characters and compound word merging. Each scenario was modeled using both LDA and BERTopic with BERTopic optimization performed through a grid search over 126 HDBSCAN hyperparameter combinations per scenario. The best model was selected by combining quantitative evaluation comprising topic coherence topic uniqueness document coverage and factuality with qualitative inspection of the intertopic distance map.

The results show that the influence of HDBSCAN hyperparameters on topic quality is not uniform across preprocessing scenarios. Among the four hyperparameters tested `cluster_selection_method` proved the most influential particularly in reducing the outlier ratio. The `eom` method consistently produced a lower outlier ratio than the `leaf` method and this difference was statistically significant based on a paired Wilcoxon signed-rank test across all scenarios. Meanwhile `cluster_selection_epsilon` showed no meaningful effect on topic coherence. The optimal hyperparameter configuration also differed across preprocessing scenarios and the best `cluster_selection_method` even reversed in the scenario applying word merging due to a shift in the embedding space.

Optimizing HDBSCAN hyperparameters through grid search increased BERTopic's topic coherence by 0.148 to 0.251 compared to the default configuration accompanied by a reduced number of topics that were easier to interpret. Optimized BERTopic also consistently achieved higher topic coherence than LDA across all scenarios while LDA outperformed BERTopic in topic

uniqueness and document coverage. Both models achieved a factuality score of 1.000 across all scenarios. Qualitatively BERTopic produced more clearly separated topics with sharper thematic boundaries while LDA covered a larger proportion of documents.

The study successfully identified six dominant themes in public discourse surrounding IPB University, namely student admissions, the emotional expressions of prospective and newly admitted students, academic life and campus identity, tuition fees, campus facilities and infrastructure, and the role of IPB University in the agriculture and food security sectors. Student admissions emerged as the most dominant theme, consistently appearing across all preprocessing scenarios, while several other themes demonstrated sensitivity to the preprocessing strategies employed. These findings provide a replicable BERTopic optimization framework for analyzing Indonesian-language social media discourse and offer valuable insights for IPB University in developing more data-driven communication and reputation management strategies.

Keywords: *BERTopic, HDBSCAN, hyperparameter, LDA, topic modeling*

@iaa.campus@ipb.ac.id

IPB University



METODE LDA DAN BERTOPIC DENGAN OPTIMASI HYPERPARAMETER HDBSCAN PADA WACANA PUBLIK IPB UNIVERSITY DI X

SABRINA ADNIN KAMILA

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Magister Statistika dan Sains Data

**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
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.



Judul Tesis : Metode LDA dan BERTopic dengan Optimasi *Hyperparameter*
HDBSCAN pada Wacana Publik IPB University di X
Nama : Sabrina Adnin Kamila
NIM : M0501241042

Disetujui oleh

Pembimbing 1:

Dr. Utami Dyah Syafitri, S.Si., M.Si.

Pembimbing 2:

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

Diketahui oleh

Ketua Program Studi:

Dr. Agus Mohamad Soleh, S.Si., M.T.
NIP. 197503151999031004

Dekan Sekolah Sains Data, Matematika, dan Informatika:

Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.
NIP. 197308241997021001

Tanggal Ujian: 30 Juli 2026

Tanggal Lulus:



@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.

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta'ala* atas segala karunia-Nya sehingga tesis ini berhasil diselesaikan. Tema yang dipilih dalam penelitian ialah pemodelan topik pada wacana publik media sosial, dengan judul "Metode LDA dan BERTopic dengan Optimasi *Hyperparameter* HDBSCAN pada Wacana Publik IPB University di X". Penelitian ini menerapkan dan mengoptimasi metode *topic modeling* untuk mengidentifikasi tema-tema dominan dalam wacana publik mengenai reputasi IPB University di media sosial X, dengan data cuitan yang dikumpulkan selama periode Agustus 2023 hingga Agustus 2025.

Terima kasih penulis ucapkan kepada komisi pembimbing, Ibu Dr. Utami Dyah Syafitri, S.Si., M.Si. dan Ibu Cici Suhaeni, S.Si., M.Si., Ph.D., yang senantiasa membimbing serta banyak memberikan ilmu, motivasi, saran, dan arahan selama penyusunan karya ilmiah ini. Ucapan terima kasih juga penulis sampaikan kepada Bapak Prof. Dr. Ir. Hari Wijayanto, M.Si. selaku penguji luar komisi, serta Bapak Prof. Dr. Ir. Made Sumertajaya, M.Si. selaku pimpinan sidang ujian tesis. Penulis juga mengucapkan terima kasih kepada seluruh dosen pengajar dan staf Program Studi Statistika dan Sains Data IPB University yang telah banyak memberikan ilmu dan bimbingan kepada penulis selama perkuliahan berlangsung.

Ungkapan terima kasih juga penulis sampaikan kepada bapak, ibu, adik, dan seluruh keluarga yang senantiasa mendoakan, memberikan dukungan, serta kasih sayang yang tidak pernah putus kepada penulis. Tidak lupa, penulis mengucapkan terima kasih kepada teman-teman Program Magister Statistika dan Sains Data Angkatan 2024 atas segala dukungan, motivasi, kebersamaan, dan bantuan selama proses perkuliahan hingga penyelesaian tesis ini.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan dan kekurangan sehingga penulis memohon maaf atas segala kekurangan tersebut. Semoga tesis ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2026

Sabrina Adnin Kamila



@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	xiii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
II TINJAUAN PUSTAKA	5
2.1 <i>Topic Modeling</i>	5
2.2 <i>Preprocessing</i> Teks	5
2.3 <i>Latent Dirichlet Allocation</i> (LDA)	6
2.4 <i>Bidirectional Encoder Representations from Transformers for Topic Modeling</i> (BERTopic)	7
2.5 <i>Grid Search</i>	10
2.6 Evaluasi	11
III METODE	14
3.1 Data	14
3.2 Prosedur Analisis Data	14
IV HASIL DAN PEMBAHASAN	19
4.1 Hasil Pengambilan dan <i>Preprocessing</i> Data	19
4.2 Hasil Pemodelan BERTopic dengan Konfigurasi <i>Default</i>	20
4.3 Hasil Optimasi <i>Hyperparameter</i> HDBSCAN	23
4.4 Pengaruh Strategi <i>Preprocessing</i> terhadap Struktur dan Kualitas Topik	33
4.5 Pengaruh <i>Hyperparameter</i> HDBSCAN terhadap Struktur Topik	34
4.6 Hasil Pemodelan LDA	38
4.7 Perbandingan Hasil Topik BERTopic Optimal dan LDA	42
4.8 Tema Dominan dalam Wacana Publik mengenai Reputasi IPB University	56
V SIMPULAN DAN SARAN	59
5.1 Simpulan	59
5.2 Saran	59
DAFTAR PUSTAKA	60
LAMPIRAN	64
RIWAYAT HIDUP	74



DAFTAR TABEL

1	Statistik data hasil <i>preprocessing</i>	19
2	Contoh transformasi teks pada setiap skenario <i>preprocessing</i>	19
3	Hasil BERTopic dengan konfigurasi <i>default</i>	20
4	Contoh topik hasil BERTopic <i>default</i> pada Skenario 1	21
5	Contoh topik hasil BERTopic <i>default</i> pada Skenario 2	21
6	Contoh topik hasil BERTopic <i>default</i> pada Skenario 3	22
7	Proporsi keragaman yang dijelaskan oleh setiap komponen utama PCA	24
8	Nilai <i>loading</i> komponen utama PCA per skenario	25
9	Konfigurasi kandidat untuk evaluasi visual beserta metrik performanya	28
10	Konfigurasi <i>hyperparameter</i> HDBSCAN terpilih per skenario	32
11	Matriks jarak Euclidean antara dokumen acuan (P0) dan sembilan tetangga terdekatnya pada ruang <i>embedding</i> hasil UMAP pada Data Skenario 1	35
12	Hasil uji <i>Wilcoxon signed-rank</i> berpasangan untuk rasio <i>outlier</i> antara metode <i>eom</i> dan <i>leaf</i>	37
13	Hasil pemodelan LDA terpilih per skenario	40
14	Perbandingan performa BERTopic teroptimasi dan LDA	42
15	Topik hasil BERTopic teroptimasi Skenario 1	44
16	Topik hasil BERTopic teroptimasi Skenario 2	46
17	Topik hasil BERTopic teroptimasi Skenario 3	49
18	Topik hasil LDA Skenario 1	51
19	Topik hasil LDA Skenario 2	52
20	Topik hasil LDA Skenario 3	54

DAFTAR GAMBAR

1	Diagram Alir Penelitian	18
2	<i>Biplot</i> PCA Skenario 1	25
3	<i>Biplot</i> PCA Skenario 2	26
4	<i>Biplot</i> PCA Skenario 3	26
5	Visualisasi <i>intertopic distance</i> pada kandidat konfigurasi HDBSCAN Skenario 1. (a) Konfigurasi [300, 5, <i>eom</i> , 0,02]. (b) Konfigurasi [300, 3, <i>eom</i> , 0,02]. (c) Konfigurasi [400, 5, <i>leaf</i> , 0,05]. (d) Konfigurasi [500, 5, <i>leaf</i> , 0,02].	29
6	Visualisasi <i>intertopic distance</i> pada kandidat konfigurasi HDBSCAN Skenario 2. (a) Konfigurasi [400, 3, <i>eom</i> , 0]. (b) Konfigurasi [300, 3, <i>eom</i> , 0,05]. (c) Konfigurasi [300, 1, <i>eom</i> , 0,02]. (d) Konfigurasi [500, 5, <i>leaf</i> , 0,02].	30
7	Visualisasi <i>intertopic distance</i> pada kandidat konfigurasi HDBSCAN Skenario 3. (a) Konfigurasi [400, 1, <i>eom</i> , 0,05]. (b) Konfigurasi [600, 3, <i>leaf</i> , 0,02]. (c) Konfigurasi [500, 1, <i>eom</i> , 0,05]. (d) Konfigurasi [400, 5, <i>leaf</i> , 0]. (e) Konfigurasi [1000, 3, <i>leaf</i> , 0].	31



8	Nilai <i>coherence</i> model LDA pada rentang $K = 1 - 10$. (a) Skenario 1. (b) Skenario 2. (c) Skenario 3.	39
9	Visualisasi <i>LDavis</i> model LDA terpilih. (a) Skenario 1. (b) Skenario 2. (c) Skenario 3.	41
10	<i>Wordcloud</i> hasil BERTopic teroptimasi Skenario 1. (a) Topik -1. (b) Topik 0. (c) Topik 1. (d) Topik 2.	45
11	<i>Wordcloud</i> hasil BERTopic teroptimasi Skenario 3. (a) Topik -1. (b) Topik 0. (c) Topik 1. (d) Topik 2. (e) Topik 3. (f) Topik 4.	51
12	<i>Wordcloud</i> hasil LDA Skenario 1. (a) Topik 1. (b) Topik 2. (c) Topik 3. (d) Topik 4.	52
13	<i>Wordcloud</i> hasil LDA Skenario 2. (a) Topik 1. (b) Topik 2. (c) Topik 3. (d) Topik 4. (e) Topik 5.	53
14	<i>Wordcloud</i> hasil LDA Skenario 3. (a) Topik 1. (b) Topik 2. (c) Topik 3. (d) Topik 4. (e) Topik 5. (f) Topik 6. (g) Topik 7. (h) Topik 8. (i) Topik 9.	55

DAFTAR LAMPIRAN

1	Lampiran 1 Hasil kombinasi lengkap <i>grid search</i> pada Skenario 1	65
2	Lampiran 2 Hasil kombinasi lengkap <i>grid search</i> pada Skenario 2	68
3	Lampiran 3 Hasil kombinasi lengkap <i>grid search</i> pada Skenario 3	71