



- 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.

PERBANDINGAN *CLUSTERING* K-MEANS DAN HDBSCAN PADA *NEURAL TOPIC MODELING* MENGGUNAKAN *EMBEDDING* INDOSBERT TERHADAP ULASAN PENGGUNA APLIKASI MYTELKOMSEL

DIVA NISFU MUSTIKA



PROGRAM STUDI STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul "Perbandingan *Clustering K-Means* Dan *Hdbscan* pada *Neural Topic Modeling* Menggunakan *Embedding IndoSBERT* Terhadap Ulasan Pengguna Aplikasi Mytelkomsel" 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, 2 Juni 2025

Diva Nisfu Mustika
G1401211002

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.



ABSTRAK

DIVA NISFU MUSTIKA. Perbandingan *Clustering* K-Means dan HDBSCAN pada *Neural Topic Modeling* Menggunakan *Embedding* IndoSBERT Terhadap Ulasan Pengguna Aplikasi Mytelkomsel. Dibimbing oleh HARI WIJAYANTO dan GERRY ALFA DITO.

Ketergantungan masyarakat terhadap layanan telekomunikasi semakin memperkuat peran penyedia akses telekomunikasi yang andal. Telkomsel, sebagai salah satu operator seluler terbesar di Indonesia, menghadirkan aplikasi MyTelkomsel untuk mempermudah pelanggan dalam mengakses berbagai layanan. Aplikasi ini telah menerima jutaan ulasan yang mencerminkan beragam pengalaman dan penilaian pengguna. Penelitian ini bertujuan untuk membandingkan kinerja algoritma K-Means dan HDBSCAN dalam BERTopic berdasarkan metrik koherensi dan keragaman topik, serta mengidentifikasi topik yang dibahas dalam ulasan pengguna MyTelkomsel. Sebanyak 88.057 ulasan dikumpulkan dari periode 1 November 2024 hingga 28 Februari 2025. Pemodelan topik dilakukan menggunakan BERTopic dengan *embedding* kalimat dari model IndoSBERT, reduksi dimensi dengan UMAP, dan penggerombolan menggunakan K-Means serta HDBSCAN. Representasi topik diperoleh melalui pembobotan c-TF-IDF dan KeyBERTInspired untuk meningkatkan kualitas kata kunci. Evaluasi menggunakan metrik koherensi dan keragaman menunjukkan bahwa konfigurasi terbaik diperoleh dari K-Means dengan representasi topik bigram dan penyertaan hiperparameter *stopwords* serta KeyBERTInspired. Konfigurasi ini menghasilkan skor koherensi sebesar 0,76 dan keragaman 0,94. Topik dominan yang teridentifikasi mencakup keluhan terkait ketidakstabilan sinyal, mahal nya paket data, masalah performa aplikasi, serta beberapa topik positif yang menyoroti kepuasan dan apresiasi terhadap fitur layanan.

Kata kunci: bertopic, k-means, hdbscan, mytelkomsel, ulasan pengguna.



ABSTRACT

DIVA NISFU MUSTIKA. Comparison of K-Means and HDBSCAN Clustering in Neural Topic Modeling Using IndoSBERT Embedding to Mytelkomsel Application User Reviews. Supervised by HARI WIJAYANTO and GERRY ALFA DITO.

The public's dependence on telecommunication services further strengthens the role of reliable telecommunication access providers. Telkomsel, one of the largest mobile operators in Indonesia, has launched the MyTelkomsel app to make it easier for customers to access various services. The app has received millions of reviews reflecting a wide range of user experiences and assessments. This study aims to compare the performance of the K-Means and HDBSCAN algorithms in BERTopic based on topic coherence and diversity metrics, as well as to identify the topics discussed in MyTelkomsel user reviews. A total of 88,057 reviews were collected from November 1, 2024, to February 28, 2025. Topic modeling was performed using BERTopic with sentence embeddings from the IndoSBERT model, dimension reduction with UMAP, and clustering using K-Means and HDBSCAN. Topic representations were obtained through c-TF-IDF and KeyBERTInspired weighting to improve keyword quality. Evaluation using coherence and diversity metrics shows that the best configuration is obtained from K-Means with bigram topic representation, inclusion of stopwords, and KeyBERTInspired. This configuration yielded a coherence score of 0.76 and a diversity of 0.94. The dominant topics identified include complaints related to signal instability, expensive data packages, application performance issues, and some positive topics highlighting satisfaction and appreciation for service features.

Keywords: bertopic, k-means, hdbscan, mytelkomsel, user reviews



- 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.

PERBANDINGAN *CLUSTERING* K-MEANS DAN HDBSCAN PADA *NEURAL TOPIC MODELING* MENGGUNAKAN *EMBEDDING* INDOSBERT TERHADAP ULASAN PENGGUNA APLIKASI MYTELKOMSEL

DIVA NISFU MUSTIKA

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

**PROGRAM STUDI STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@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 Skripsi : Perbandingan *Clustering* K-Means dan HDBSCAN pada *Neural Topic Modeling* Menggunakan *Embedding* IndoSBERT Terhadap Ulasan Pengguna Aplikasi MyTelkomsel

Nama : Diva Nisfu Mustika
NIM : G1401211002

Disetujui oleh

Pembimbing 1:
Prof. Dr. Ir. Hari Wijayanto, M.Si



Pembimbing 2:
Gerry Alfa Dito, S.Si, M.Si



Diketahui oleh

Ketua Program Studi:
Dr. Bagus Sartono, S.Si., M.Si
NIP 19780411 200501 1 002



Tanggal Ujian:
10 Juni 2025

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Judul yang dipilih dalam penelitian ini "Perbandingan *Clustering K-Means* dan *HDBSCAN* pada *Neural Topic Modeling* Menggunakan *Embedding IndoSBERT* Terhadap Ulasan Pengguna Aplikasi Mytelkomsel". Pada kesempatan ini, penulis ingin mengucapkan terima kasih kepada seluruh pihak yang telah berperan sejak penulis menempuh studi sampai proses penulisan karya ilmiah ini berakhir, di antaranya:

1. Bapak Ahmad Ya'i dan Ibu Lindung Sari selaku orang tua, Wildan Abdillah selaku saudara kandung, serta seluruh keluarga besar yang telah memberi restu, doa, dukungan, dan motivasinya.
2. Bapak Prof. Dr. Ir. Hari Wijayanto, M.Si dan Bapak Gerry Alfa Dito, S.Si, M.Si selaku komisi pembimbing skripsi yang telah memberikan banyak arahan dan saran dalam penulisan karya ilmiah ini.
3. Bapak La Ode Abdul Rahman, S.Si., M.Si selaku moderator kolokium, Ibu Pika Silvianti, S.Si., M.Si selaku moderator seminar dan Prof. Dr. Ir. Muhammad Nur Aidi, M.S selaku dosen penguji yang telah memberikan kritik serta masukan untuk perbaikan karya ilmiah ini.
4. Seluruh dosen dan tenaga kependidikan Program Studi Statistika dan Sains Data IPB yang telah menyampaikan ilmunya dan membantu selama proses perkuliahan hingga menyelesaikan studi.
5. Teman-teman dari Statistika 58 khususnya Andi Fatihatul Fuadiyah Akbar, Angel Martha Pradina Pangaribuan, Dhiya Khalishah Tsany Suwarso, Kheni Hikmah Lestari, Reyza Siva Dewi yang telah menemani, membantu, memberi banyak saran, masukan, motivasi, dukungan, serta semangat selama masa kuliah ini.
6. Beta Maisaroh selaku teman penulis yang telah menjadi rekan baik dan penyemangat dari masa SMP hingga sekarang.
7. Teman-teman dari departemen analisis data GSB 2023, tim BDC Satria Data 2024 khususnya Dzakwan Alifi serta teman-teman dari departemen data analytics GSB 2024 khususnya Azizah Amalia Azra, Hanifah Rahmacindia Nasution, Lutfi Syahreza Lubis, Much Fazrin Sepranjani Fatah, dan Nabil Naufal yang juga telah menemani, membantu, dan selalu menyemangati selama masa kuliah.
8. Kamilah Nurul Azizah, Asfiah Adiba, Oktavia Galih Pratiwi, dan Arfiah Kania Sektiaruni yang telah menjadi teman diskusi hingga penyemangat dalam penulisan karya ilmiah ini.

Penulis menyadari bahwa karya ilmiah ini masih belum mencapai tingkat kesempurnaan yang diharapkan. Oleh karena itu, penulis memohon maaf apabila ditemukan kesalahan dan kekurangan dalam karya ilmiah ini. Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan berkontribusi pada kemajuan ilmu pengetahuan.

Bogor, 2 Juni 2025

Diva Nisfu Mustika

DAFTAR ISI

DAFTAR TABEL	ix
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	xi
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	3
II TINJAUAN PUSTAKA	4
2.1 MyTelkomsel	4
2.2 <i>Web Scraping</i>	4
2.3 <i>Neural Topic Modeling</i>	5
2.4 BERTopic	5
2.5 IndoSBERT	6
2.6 <i>Uniform Manifold Approximation and Projection (UMAP)</i>	9
2.7 K-Means	12
2.8 <i>Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN)</i>	12
2.9 Tokenisasi N-gram	15
2.10 <i>Bag-of-Words (BoW)</i>	15
2.11 <i>Class Based Term Frequency Inverse Document Frequency (c-TF-IDF)</i>	16
2.12 KeyBERTInspired	16
2.13 ChatGPT	17
2.14 Koherensi topik	17
2.15 Keragaman Topik	18
III METODE	19
3.1 Data	19
3.2 Prosedur Analisis Data	19
IV HASIL DAN PEMBAHASAN	22
4.1 Prapemrosesan Data	22
4.2 Pemodelan topik dengan BERTopic	22
4.2.1 <i>Embedding</i> teks	22
4.2.2 Reduksi dimensi dan penggerombolan	23
4.2.3 Representasi topik	23
4.3 Evaluasi Pemodelan Topik dengan BERTopic	24
4.3.1 Pemodelan Topik K-Means dengan Representasi Topik Unigram... ..	24
4.3.2 Pemodelan Topik K-Means dengan Representasi Topik Unigram	26
4.3.3 Pemodelan Topik K-Means dengan Representasi Topik Bigram	28
4.3.4 Pemodelan Topik HDBSCAN dengan Representasi Topik Unigram	30
4.3.5 Pemodelan Topik HDBSCAN dengan Representasi Topik Unigram	32

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.

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.

4.3.6	Pemodelan Topik HDBSCAN dengan Representasi Topik Bigram .	34
4.4	Perbandingan Hasil Evaluasi	36
4.5	Pemilihan Konfigurasi Pemodelan Topik Terbaik	38
4.6	Interpretasi Pemodelan Topik Terbaik	40
	SIMPULAN DAN SARAN	42
5.1	Simpulan	42
5.2	Saran	42
	DAFTAR PUSTAKA.....	43
	LAMPIRAN	47
	RIWAYAT HIDUP	56



DAFTAR TABEL

1	Rincian peubah hasil <i>scraping</i>	19
2	Nilai hiperparameter	20
3	Contoh data sebelum dan sesudah prapemrosesan	22
4	Contoh data sebelum dan sesudah pembentukan <i>embedding</i>	22
5	Contoh data sebelum dan sesudah reduksi dimensi	23
6	Contoh hasil tokenisasi unigram, unigram, dan bigram	24
7	Contoh hasil pembobotan c-tf-idf unigram (1,1)	24
8	Tiga topik dominan hasil pemodelan topik k-means topik unigram	26
9	Tiga topik hasil pemodelan topik k-means topik unigram	28
10	Tiga topik dominan hasil pemodelan topik k-means topik bigram	30
11	Tiga topik dominan hasil pemodelan topik hdbscan topik unigram	32
12	Tiga topik dominan hasil pemodelan topik hdbscan topik unigram	34
13	Tiga topik dominan hasil pemodelan topik hdbscan topik bigram	36
14	Hasil evaluasi pemodelan topik awal dan <i>tuning</i> hiperparameter	37
15	Hasil evaluasi pemodelan topik periode 1 dan 2	39



DAFTAR GAMBAR

1	Tahapan pemodelan topik menggunakan bertopic	5
2	Arsitektur indosbert (Diana dan Khodra 2023)	7
3	Representasi input bert (Devlin <i>et al.</i> 2019)	7
4	Ilustrasi reduksi dimensi dengan umap (Oide dan Sugita 2022)	10
5	Ilustrasi <i>core distance</i> dalam hdbscan (McInness <i>et al.</i> 2016)	13
6	Contoh struktur mst (McInness <i>et al.</i> 2016)	13
7	Contoh hierarki gerombol (McInness <i>et al.</i> 2016)	14
8	Contoh <i>condensed tree</i> (McInness <i>et al.</i> 2016)	14
9	Algoritma keybertinspired (Grootendorst 2022)	16
10	Awan kata tiga topik dominan k-means topik unigram sebelum <i>tuning</i> representasi topik	25
11	Awan kata tiga topik dominan k-means topik unigram sesudah <i>tuning</i> representasi topik	25
12	Awan kata tiga topik k-means topik unigram sebelum <i>tuning</i> representasi topik	27
13	Awan kata tiga topik k-means topik unigram sesudah <i>tuning</i> representasi topik	27
14	Awan kata tiga topik dominan k-means topik bigram sebelum <i>tuning</i> representasi topik	29
15	Awan kata tiga topik dominan k-means topik bigram setelah <i>tuning</i> representasi topik	29
16	Awan kata tiga topik dominan hdbscan topik unigram sebelum <i>tuning</i> representasi topik	31
17	Awan kata tiga topik dominan hdbscan topik unigram sesudah <i>tuning</i> representasi topik	31
18	Awan kata tiga topik dominan hdbscan topik unigram sebelum <i>tuning</i> representasi topik	33
19	Awan kata tiga topik dominan hdbscan topik unigram sebelum <i>tuning</i> representasi topik	33
20	Awan kata tiga topik dominan hdbscan topik bigram sebelum <i>tuning</i> representasi topik	35
21	Awan kata tiga topik dominan hdbscan topik bigram sesudah <i>tuning</i> representasi topik	35
22	Visualisasi hdbscan topik bigram	38
23	Visualisasi k-means topik bigram	39

DAFTAR LAMPIRAN

1	Lampiran 1 Hasil evaluasi seluruh kombinasi hiperparameter k-means topik unigram	48
2	Lampiran 2 Hasil evaluasi seluruh kombinasi hiperparameter k-means topik unigram	49
3	Lampiran 3 Hasil evaluasi seluruh kombinasi hiperparameter k-means topik bigram	50
4	Lampiran 4 Hasil evaluasi seluruh kombinasi hiperparameter hdbscan topik unigram	51
5	Lampiran 5 Hasil evaluasi seluruh kombinasi hiperparameter hdbscan topik unigram	52
6	Lampiran 6 Hasil evaluasi seluruh kombinasi hiperparameter hdbscan topik bigram	53
7	Lampiran 7 Hasil pemodelan topik konfigurasi terbaik	54



- 