

- 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.

***PROBABILISTIC TOPIC MODELING* DAN PERBANDINGAN ALGORITMA ANALISIS SENTIMEN DENGAN PENDEKATAN *HYBRID* PADA ULASAN WARDAH GLASTING LIQUID LIP**

SABRINA



**PROGRAM STUDI MATEMATIKA
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 “*Probabilistic Topic Modeling* dan Perbandingan Algoritma Analisis Sentimen dengan Pendekatan *Hybrid* pada Ulasan Wardah Glasting Liquid Lip” 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, Juli 2025

Sabrina
G5401211072

ABSTRAK

SABRINA. *Probabilistic Topic Modeling* dan Perbandingan Algoritma Analisis Sentimen dengan Pendekatan *Hybrid* pada Ulasan Wardah Glasting Liquid Lip. Dibimbing oleh MOCHAMAD TITO JULIANTO dan MOHAMAD KHOIRUN NAJIB.

Wardah menjadi merek lokal dengan penjualan lipstick terbanyak di Shopee pada Juni 2024. Salah satu produknya yaitu Wardah Glasting Liquid Lip. Penelitian ini bertujuan mengidentifikasi topik utama pada ulasan Wardah Glasting Liquid Lip di Shopee menggunakan metode *Latent Dirichlet Allocation* dan membandingkan performa tiga algoritma analisis sentimen yaitu *Support Vector Machine*, *CatBoost Classifier*, dan *Long Short-Term Memory* dengan pendekatan *hybrid*. Perbandingan didasarkan pada *accuracy*, *precision*, *recall*, dan *F1-score*. Pelabelan sentimen dilakukan secara manual dan *word embedding* menggunakan IndoBERT. Proses *topic modeling* menghasilkan tiga topik utama yaitu tekstur, pengiriman, dan warna. Topik dengan sentimen positif terbanyak yaitu warna dan sentimen negatif terbanyak yaitu tekstur. Model *Long Short-Term Memory* menjadi model terbaik dengan *accuracy* sebesar 73,2%. Akan tetapi, ketiga model masih mengalami kesulitan dalam mengklasifikasikan sentimen netral dengan rendahnya nilai *F1-score* pada kelas tersebut yang hanya berada di rentang 14-29%.

Kata kunci: analisis sentimen, *catboost classssifier*, *latent dirichlet allocation*, *long short-term memory*, *support vector machine*

ABSTRACT

SABRINA. *Probabilistic Topic Modeling and Comparison of Sentiment Analysis Algorithms Using Hybrid Approaches for Wardah Glasting Liquid Lip*. Supervised by MOCHAMAD TITO JULIANTO and MOHAMAD KHOIRUN NAJIB.

Wardah became the local brand with the highest lipstick sales on Shopee in June 2024. One of its products is the Wardah Glasting Liquid Lip. This study aims to identify the main topics in reviews of the Wardah Glasting Liquid Lip on Shopee using the *Latent Dirichlet Allocation* method and to compare the performance of three sentiment analysis algorithms: *Support Vector Machine*, *CatBoost Classifier*, and *Long Short-Term Memory* using a *hybrid approach*. The comparison is based on *accuracy*, *precision*, *recall*, and *F1-score*. Sentiment labeling was performed manually, while word embedding utilized IndoBERT. The topic modeling process revealed three predominant topics: texture, shipping, and color. The topic with the most positive sentiment was color, while the most negative sentiment was texture. The *Long Short-Term Memory* model performed best, achieving an accuracy of 73,2%. However, all three models struggled to classify neutral sentiment, as indicated by the low *F1-score* in that class which ranged from just 14% to 29%..

Keywords: *catboost classssifier*, *latent dirichlet allocation*, *long short-term memory*, *sentiment analysis*, *support vector machine*



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 2025
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.

***PROBABILISTIC TOPIC MODELING* DAN PERBANDINGAN
ALGORITMA ANALISIS SENTIMEN DENGAN
PENDEKATAN *HYBRID* PADA ULASAN WARDAH
GLASTING LIQUID LIP**

SABRINA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Matematika pada
Program Studi Matematika

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



@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.

Tim Penguji pada Ujian Skripsi:
Prof. Dr. Ir. Sri Nurdianti, M.Sc.

Judul Skripsi : *Probabilistic Topic Modeling* dan Perbandingan Algoritma Analisis Sentimen dengan Pendekatan *Hybrid* pada Ulasan Wardah Glasting Liquid Lip

Nama : Sabrina
NIM : G5401211072

Disetujui oleh

Pembimbing 1:
Mochamad Tito Julianto, S.Si., M.Kom.

Pembimbing 2:
Mohamad Khoirun Najib, S.Si., M.Mat.



Diketahui oleh

Ketua Program Studi:
Dr. Donny Citra Lesmana, M.Fin.Math.
NIP 197902272005011001



Tanggal Ujian: 18 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. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Desember 2024 sampai bulan April 2025 ini ialah analisis data ulasan produk di Shopee dengan judul *Probabilistic Topic Modeling* dan Perbandingan Algoritma Analisis Sentimen dengan Pendekatan *Hybrid* pada Ulasan Wardah Glasting Liquid Lip". Terima kasih penulis ucapkan kepada pihak-pihak yang telah memberikan dukungan dan arahan selama proses penulisan skripsi berlangsung, di antaranya:

1. Ibu Emah Yati, Tante Yeti Susilawati, dan Maudy Tegar Utami selaku ibu, tante, dan sepupu dari penulis yang telah memberikan dukungan, semangat, motivasi, kasih sayang, dan do'a kepada penulis sampai hari ini.
2. Bapak Mochamad Tito Julianto, S.Si., M.Kom. selaku dosen pembimbing I, Bapak Mohamad Khoirun Najib, S.Si., M.Mat. selaku dosen pembimbing II, dan Ibu Prof. Dr. Ir. Sri Nurdianti, M.Sc. selaku dosen penguji yang telah memberikan ilmu, masukan, dan saran kepada penulis dalam penulisan dan perbaikan karya ilmiah ini.
3. Bapak Yunus dan Ibu Mami selaku ayah dan keluarga penulis yang telah memberikan do'a kepada penulis.
4. Seluruh dosen, asisten dosen, dan staf di Program Studi Matematika SSMI IPB yang telah memberikan ilmu kepada penulis dan membantu dalam proses administrasi.
5. Kak Lidya Lorenta selaku mentor penulis saat magang yang telah memberikan ilmu sehingga muncul inspirasi ide untuk penulisan karya ilmiah ini.
6. Aida Zaskia Febrianty, Salsabila Yonita, Anisa Nurhajiza, dan Shiva Ananda selaku sahabat terdekat di Program Studi Matematika SSMI IPB yang selalu menemani dan memberikan dukungan serta semangat kepada penulis.
7. Sylvia Wardani selaku sahabat terdekat penulis sejak PKU sampai sekarang yang selalu menemani, membantu, dan mendukung penulis.
8. Rusmiyati, Annisa Nur Aulia Sandy, dan Putri Amelia Silvana selaku sahabat penulis sejak SMA sampai sekarang yang selalu mendukung dan mendo'akan penulis.
9. Grup Homomorfisma, teman satu bimbingan penulis, dan seluruh teman-teman matematika angkatan 58 yang telah membantu penulis selama perkuliahan.
10. Serta seluruh pihak yang telah membantu penulis selama perkuliahan yang tidak dapat disebutkan satu-satu.

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

Bogor, Juli 2025

Sabrina

DAFTAR TABEL	viii
DAFTAR GAMBAR	viii
DAFTAR LAMPIRAN	ix
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	3
II TINJAUAN PUSTAKA	4
2.1 Analisis Sentimen	4
2.2 Rumus Slovin	4
2.3 <i>Topic Modeling</i>	4
2.4 <i>Topic Coherence</i>	5
2.5 Model <i>Cv</i> sebagai Model <i>Coherence Measure</i>	7
2.6 <i>Bag of Words</i> (BoW)	8
2.7 <i>Latent Dirichlet allocation</i> (LDA)	9
2.8 <i>Support Vector Machine</i> (SVM)	9
2.9 <i>CatBoost Classifier</i>	11
2.10 <i>Long Short-Term Memory</i> (LSTM)	11
2.11 <i>Cross-Entropy Loss Function</i>	13
2.12 <i>Adam Optimizer</i>	13
2.13 IndoBERT	14
2.14 Evaluasi Model Klasifikasi	15
III METODE	18
3.1 Sumber dan Jenis Data	18
3.2 Tahapan Penelitian	19
3.2.1 <i>Pre-processing</i> data	19
3.2.2 <i>Topic modeling</i>	21
3.2.3 Pelabelan sentimen	22
3.2.4 Pembagian data	22
3.2.5 <i>Word embedding</i> dan ekstraksi fitur dengan IndoBERT	22
3.2.6 Pemodelan klasifikasi sentimen	22
3.2.7 Evaluasi dan perbandingan model	23
3.3 Lingkungan Pengembangan	23
3.3.1 Perangkat keras	24
3.3.2 Perangkat lunak	24
IV HASIL DAN PEMBAHASAN	26
4.1 <i>Pre-processing</i> Data	26
4.2 <i>Topic Modeling</i>	29
4.3 Pelabelan Sentimen	34
4.4 Eksplorasi Data	35
4.5 Pembagian Data	40
4.6 <i>Word Embedding</i> dengan IndoBERT	41
4.7 Model Klasifikasi Sentimen	42
4.7.1 <i>Support Vector Machine</i>	42

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 mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



4.7.2 <i>CatBoost Classifier</i>	43
4.7.3 <i>Long Short-Term Memory</i>	44
4.8 Evaluasi dan Perbandingan Model Klasifikasi	47
V SIMPULAN DAN SARAN	49
5.1 Simpulan	49
5.2 Saran	49
DAFTAR PUSTAKA	50
LAMPIRAN	55

DAFTAR TABEL

1	<i>Confusion matrix</i>	15
2	Jumlah total ulasan dan banyaknya ulasan yang diambil dari tiap <i>rating</i>	18
3	Contoh ulasan untuk masing-masing <i>rating</i>	19
4	Nilai parameter model <i>long short-term memory</i>	23
5	Hasil <i>case folding</i> pada contoh data	26
6	Hasil <i>data cleaning</i> pada contoh data	26
7	Hasil <i>tokenization</i> pada contoh data	27
8	Contoh daftar kata <i>slang</i> pada kamus kata acuan	27
9	Hasil <i>normalization</i> pada contoh data	27
10	Hasil <i>translation</i> pada contoh data	28
11	Hasil <i>stopword removal</i> pada contoh data	28
12	Hasil <i>stemming</i> pada contoh data	28
13	Hasil vektorisasi <i>bag of words</i> pada contoh data	29
14	Nilai <i>coherence score</i> untuk tiap percobaan	30
15	Hasil <i>topic modeling</i> dengan <i>latent dirichlet allocation</i> dan interpretasinya	31
16	Contoh hasil <i>topic labelling</i>	33
17	Contoh hasil pelabelan sentimen	35
18	Rincian pembagian data	41
19	Bentuk ukuran data peubah bebas hasil ekstraksi fitur untuk tiap algoritma	42
20	Nilai parameter untuk model percobaan <i>support vector machine</i>	42
21	Nilai parameter untuk model percobaan <i>catboost classifier</i>	44
22	Arsitektur model percobaan <i>long short-term memory</i>	45
23	Nilai <i>performance metrics</i> untuk ketiga model	47

DAFTAR GAMBAR

1	Struktur umum <i>topic coherence</i> (Röder et al. 2015)	5
2	Contoh penentuan <i>topic coherence</i> dari <i>topic</i> {"cat", "dog", "toy"} (Pedro 2022)	6
3	Ilustrasi <i>support vector machine</i> (Kecman 2005)	10

4	Ilustrasi pendekatan <i>one-against-all</i> (Andrew 2024)	10
5	Arsitektur LSTM (Zhang <i>et al.</i> 2018)	12
6	Contoh <i>word embedding</i> dengan IndoBERT (Imron <i>et al.</i> 2023)	15
7	Diagram alir penelitian	25
8	Grafik nilai <i>coherence score</i>	30
9	Hasil visualisasi LDA: (a) topik 1; (b) topik 2; dan (c) topik 3	32
10	Visualisasi jumlah ulasan untuk tiap label topik	34
11	Visualisasi jumlah ulasan untuk tiap label sentimen	34
12	Visualisasi jumlah ulasan berdasarkan bulan di tahun 2024	36
13	Visualisasi jumlah ulasan berdasarkan sentimen untuk tiap topik pada bulan Januari-Desember di tahun 2024	37
14	Visualisasi jumlah ulasan berdasarkan topik untuk tiap sentimen di tahun 2024	38
15	Visualisasi awan kata ulasan sentimen positif	39
16	Visualisasi awan kata ulasan sentimen negatif	40
17	Visualisasi awan kata ulasan sentimen netral	40
18	Hasil tokenisasi <i>word embedding</i> dengan IndoBERT: (a) data latih; (b) data validasi; dan (c) data uji	41
19	<i>Confusion matrix</i> model <i>support vector machine</i>	43
20	<i>Confusion matrix</i> model <i>catboost classifier</i>	44
21	<i>Plot</i> model arsitektur <i>long short-term memory</i> terpilih	45
22	Grafik nilai <i>validation loss</i> dan <i>training loss</i> serta grafik <i>validation accuracy</i> dan <i>training accuracy</i>	46
23	<i>Confusion matrix</i> model <i>long short-term memory</i>	47

DAFTAR LAMPIRAN

1	Lampiran 1 <i>Code/syntax</i> yang dijalankan selama proses penelitian	56
2	Lampiran 2 Hasil visualisasi LDA topik 2: (a) peta jarak antar topik dan (b) 15 kata teratas (paling relevan) dari topik 2	67
3	Lampiran 3 Hasil visualisasi LDA topik 3: (a) peta jarak antar topik dan (b) 15 kata teratas (paling relevan) dari topik 3	68