

PENERAPAN *MULTINOMIAL NAÏVE BAYES* DAN LIGHTGBM DALAM KLASIFIKASI SENTIMEN PADA ISU *OVERCLAIM SKINCARE*

ALINKA NAJWA SONIA



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



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Penerapan *Multinomial Naïve Bayes* dan LightGBM dalam Klasifikasi Sentimen pada Isu *Overclaim Skincare*” 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 2026

Alinka Najwa Sonia
NIM. G5401221067

ABSTRAK

ALINKA NAJWA SONIA. Penerapan *Multinomial Naïve Bayes* dan *LightGBM* dalam Klasifikasi Sentimen pada Isu *Overclaim Skincare*. Dibimbing oleh ELIS KHATIZAH dan MOHAMAD KHOIRUN NAJIB.

Fenomena *overclaim skincare* di TikTok memunculkan berbagai respons publik yang tercermin pada kolom komentar sehingga penting untuk memahami sentimen yang terbentuk. Penelitian ini menerapkan model *Multinomial Naïve Bayes* (MNB) dan *Light Gradient Boosting Machine* (LightGBM) untuk mengklasifikasikan sentimen komentar ke dalam tiga kelas. Selain itu, penelitian ini juga menilai pengaruh jenis pelabelan dan distribusi sentimen terhadap performa model, serta mengidentifikasi kata-kata pemicu sentimen. Ketidakseimbangan distribusi data diatasi melalui augmentasi menggunakan teknik *back translation*. Model diuji pada empat skenario berdasarkan jenis pelabelan dan kondisi distribusi sentimen, lalu dievaluasi dengan akurasi, presisi, *recall*, dan *f1-score*. MNB pada skenario pelabelan manual dengan distribusi sentimen seimbang ditemukan sebagai model terbaik dengan akurasi 0,81 dan performa lebih stabil dibandingkan LightGBM. Temuan ini menegaskan bahwa kualitas pelabelan dan keseimbangan distribusi sentimen berperan krusial dalam meningkatkan kinerja model. Selain itu, kata-kata pemicu sentimen berhasil diidentifikasi melalui analisis *feature importance* berbasis TF-IDF pada model terbaik.

Kata kunci: analisis sentimen, LightGBM, *Multinomial Naïve Bayes*, *overclaim skincare*, tiktok

ABSTRACT

ALINKA NAJWA SONIA. Implementation of Multinomial Naïve Bayes and LightGBM for Sentiment Classification on Overclaim Skincare Issues. Supervised by ELIS KHATIZAH and MOHAMAD KHOIRUN NAJIB.

The discussion of overclaim skincare on TikTok has generated various public responses in the comment section, making it important to understand the sentiment formed. This study applies Multinomial Naïve Bayes (MNB) and Light Gradient Boosting Machine (LightGBM) to classify comment sentiment into three classes. This study also examines the effect of labeling methods and sentiment distribution on model performance, as well as identifying sentiment trigger words. The imbalanced dataset was addressed through back translation augmentation. The models are tested in four scenarios based on labeling methods and sentiment distribution condition, then evaluated using accuracy, precision, recall, and f1-score. MNB with manual labeling and balanced sentiment distribution was found as the best-performing model, achieving an accuracy of 0.81 and more stable performance compared to LightGBM. These findings confirm that labeling quality and balanced sentiment distribution are crucial in improving model performance. Furthermore, sentiment trigger words were successfully identified through TF-IDF-based feature importance analysis on the best model.

Keywords: LightGBM, Multinomial Naïve Bayes, overclaim skincare, sentiment analysis, tiktok



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.

PENERAPAN *MULTINOMIAL NAÏVE BAYES* DAN LIGHTGBM DALAM KLASIFIKASI SENTIMEN PADA ISU *OVERCLAIM SKINCARE*

ALINKA NAJWA SONIA

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
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.

Tim Penguji pada Ujian Skripsi:

1 Prof. Dr. Ir. Sri Nurdianti, M.Sc.



Judul Skripsi : Penerapan *Multinomial Naïve Bayes* dan *LightGBM* dalam
Klasifikasi Sentimen pada Isu *Overclaim Skincare*

Nama : Alinka Najwa Sonia
NIM : G5401221067

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Elis Khatizah, S.Si., M.Si., Ph.D.

Pembimbing 2:

Mohamad Khoirun Najib, S.Si., M.Mat.

Diketahui oleh

Ketua Program Studi:

Dr. Donny Citra Lesmana, S.Si., M.Fin.Math.
NIP. 197902272005011001

Tanggal Ujian:
3 Juni 2026

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 September 2025 sampai bulan April 2026 ini ialah klasifikasi sentimen, dengan judul “Penerapan *Multinomial Naïve Bayes* dan *LightGBM* dalam Klasifikasi Sentimen pada Isu *Overclaim Skincare*”. Penyelesaian karya ilmiah ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih yang tulus kepada:

1. Ibu Qoidah, Inesia Melati, dan Muhamad Ken Candra selaku keluarga yang sangat penulis sayangi, atas doa yang tak pernah putus dan dukungan penuhnya sehingga penulis dapat menyelesaikan studi Matematika di IPB University. Secara khusus, penulis juga menyampaikan rasa rindu dan terima kasih yang amat dalam kepada Alm. Bapak Agus Najib atas dukungan luar biasa semasa hidup, serta doa-doa yang penulis yakini senantiasa mengiringi langkah penulis dari surga,
2. Ibu Elis Khatizah, S.Si., M.Si., Ph.D. dan Bapak Mohamad Khoirun Najib, S.Si., M.Mat. selaku dosen pembimbing yang senantiasa memberikan bimbingan, motivasi, semangat, dan nasihat dengan penuh kesabaran, serta Ibu Prof. Dr. Ir. Sri Nurdianti, M.Sc. selaku dosen penguji atas segala saran dan masukan yang diberikan sehingga skripsi ini dapat diselesaikan dengan baik dan tepat waktu,
3. Seluruh dosen dan asisten dosen di Program Studi S-1 Matematika IPB University yang telah membekali penulis dengan ilmu yang sangat bermanfaat, serta seluruh staf tenaga kependidikan yang telah banyak membantu dan memperlancar segala urusan administrasi penulis selama masa perkuliahan,
4. Vidia, Alya, Zalvi, Amanda, Keris, Gandhis, Salwa, Ayuni, Shava, Hanifa, dan Griselda, yang selalu menjadi rumah yang hangat bagi penulis,
5. Teman-teman Cemara IPB, Medbrand Gantari Arti, Medbrenk Niscala Aletta, Medbrenk Akasha Phivana, Capdjay 61, Bebes Sukajaya, dan Matematika 59, yang telah memberikan banyak pengalaman berharga, cerita, serta kesan positif yang sangat mewarnai masa perkuliahan penulis,
6. Seluruh pihak yang telah memberikan doa, dukungan, dan semangat kepada penulis yang tidak bisa disebutkan satu per satu, namun keterlibatannya sangat berarti dalam penyelesaian karya ilmiah ini.

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

Bogor, Juli 2026

Alinka Najwa Sonia



DAFTAR ISI

DAFTAR TABEL	vii
DAFTAR GAMBAR	vii
DAFTAR LAMPIRAN	viii
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
TINJAUAN PUSTAKA	3
2.1 Isu <i>Overclaim Skincare</i> pada Sosial Media TikTok	3
2.2 <i>Text Mining</i> dan <i>Natural Language Processing</i>	3
2.3 Analisis Sentimen dan Klasifikasi Sentimen	3
2.4 Pra-pemrosesan dan Segmentasi Teks	4
2.5 IndoBERT (<i>Bidirectional Encoder Representations from Transformers</i>)	4
2.6 <i>Fleiss' Kappa</i>	6
2.7 <i>Weighted Majority Voting</i>	8
2.8 Penanganan Data Tak Seimbang dengan Augmentasi Data Teks	8
2.9 TF-IDF (<i>Term Frequency-Inverse Document Frequency</i>)	9
2.10 Teorema Bayes	10
2.11 <i>Naïve Bayes Classifier</i>	11
2.12 <i>Multinomial Naïve Bayes</i>	12
2.13 Estimasi Parameter dengan <i>Maximum Likelihood Estimation</i>	14
2.14 Penanganan Nilai Probabilitas Nol dengan <i>Laplace Smoothing</i>	15
2.15 <i>Gradient Boosting Decision Tree</i>	16
2.16 LightGBM (<i>Light Gradient Boosting Machine</i>)	18
2.17 Evaluasi Model	19
III METODE	22
3.1 Jenis dan Sumber Data	22
3.2 Tahapan Penelitian	22
IV HASIL DAN PEMBAHASAN	27
4.1 Pra-pemrosesan Data untuk Pelabelan Berbasis IndoBERT (Tahap A)	27
4.2 Eksplorasi Data	31
4.3 Pra-pemrosesan Data untuk Pemodelan (Tahap B)	35
4.4 Implementasi Model	36
4.5 Analisis Fitur Penting pada Model Terbaik	42
V SIMPULAN DAN SARAN	44
5.1 Simpulan	44
5.2 Saran	44
DAFTAR PUSTAKA	45
LAMPIRAN	47
RIWAYAT HIDUP	54

DAFTAR TABEL

1	Interpretasi tingkat kesepakatan nilai kappa (κ)	8
2	<i>Confusion matrix</i> untuk multikelas	20
3	Contoh pengelompokkan emosi pada representasi emoji	27
4	Contoh kalimat komentar hasil pra-pemrosesan Tahap A	28
5	Nilai <i>logit</i> setiap kelas sentimen pada sampel data pertama penelitian	28
6	Contoh perhitungan dan penentuan label akhir pada pelabelan IndoBERT	29
7	Contoh hasil pelabelan manual oleh tiga anotator pada data penelitian	29
8	Contoh perhitungan <i>agreement</i> setiap komentar pada data penelitian	30
9	Hasil perhitungan proporsi setiap kelas pada data penelitian	30
10	Contoh perhitungan dan penentuan label akhir pada pelabelan manual	31
11	Distribusi ketidaksesuaian antara label manual dan label IndoBERT	32
12	Contoh ketidaksesuaian pelabelan antara IndoBERT dan manual	32
13	Contoh hasil proses augmentasi data dengan <i>back translation</i>	34
14	Contoh kalimat komentar hasil pra-pemrosesan Tahap B	35
15	Contoh perhitungan bobot TF-IDF pada data penelitian	36
16	Hasil evaluasi performa model pada Skenario (1)	37
17	Hasil evaluasi performa model pada Skenario (2)	39
18	Hasil evaluasi performa model pada Skenario (3)	40
19	Hasil evaluasi performa model pada Skenario (4)	41
20	Kata-kata pemicu pada setiap kelas berdasarkan bobot TF-IDF	42

DAFTAR GAMBAR

1	Representasi input pada BERT	5
2	Ilustrasi alur kerja BERT	6
3	Ilustrasi <i>Bag of Words</i>	9
4	Proses kerja <i>bagging</i> dan <i>boosting</i> dalam <i>ensemble learning</i>	16
5	Perbandingan skema pembelahan <i>leaf-wise</i> dan <i>level-wise</i>	18
6	Skema <i>Gradient-base One-side Sampling</i> (GOSS) dan <i>Exclusive Feature Bundling</i> (EFB)	19
7	Tampilan Apify (TikTok Comments Scraper)	22
8	Bagan alur tahapan penelitian	26
9	Distribusi kelas sebelum dilakukan augmentasi dengan (A) pelabelan IndoBERT dan (B) pelabelan manual	33
10	Distribusi kelas setelah dilakukan augmentasi dengan (A) pelabelan IndoBERT dan (B) pelabelan manual	34
11	<i>Confusion matrix</i> pada Skenario (1) dengan model (A) <i>Multinomial Naïve Bayes</i> dan (B) <i>Light Gradient Boosting Machine</i>	37
12	<i>Confusion matrix</i> pada Skenario (2) dengan model (A) <i>Multinomial Naïve Bayes</i> dan (B) <i>Light Gradient Boosting Machine</i>	38
13	<i>Confusion matrix</i> pada Skenario (3) dengan model (A) <i>Multinomial Naïve Bayes</i> dan (B) <i>Light Gradient Boosting Machine</i>	39

DAFTAR LAMPIRAN

1	Data komentar pada konten TikTok dengan isu <i>overclaim skincare</i>	48
2	Perhitungan <i>Fleiss' Kappa</i> dan <i>weighted majority voting</i> pada pelabelan manual	48
3	Code untuk proses <i>case folding</i>	48
4	Code untuk proses representasi emoji	48
5	Code untuk proses <i>data cleaning</i>	49
6	Code untuk proses pelabelan IndoBERT	49
7	Code untuk proses pra-pemrosesan untuk pemodelan Tahap B (tokenisasi, normalisasi, <i>stopword removal</i> , <i>stemming</i>)	50
8	Code untuk proses pembobotan TF-IDF	50
9	Code untuk proses pembagian <i>data training</i> dan <i>data testing</i>	50
10	Code untuk proses augmentasi data <i>back translation</i>	51
11	Code untuk proses implementasi model <i>Multinomial Naïve Bayes</i>	52
12	Code untuk proses implementasi model LightGBM	53