



- 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 *AUTOMATIC LABELLING* DALAM KLASIFIKASI *NAIVE BAYES* UNTUK ANALISIS SENTIMEN DATA TRANSKRIPSI VIDEO PENDEK

PANDU PERSADA TANJUNG



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



@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



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan *Automatic Labelling* dalam Klasifikasi Naive Bayes untuk Analisis Sentimen Data Transkripsi Video Pendek” 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

Pandu Persada Tanjung
G6401221097

- 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.

ABSTRAK

PANDU PERSADA TANJUNG. Perbandingan *Automatic Labelling* dalam Klasifikasi Naive Bayes untuk Analisis Sentimen Data Transkripsi Video Pendek. Dibimbing oleh YANI NURHADRYANI dan HAFIDLOTUL FATIMAH AHMAD.

Pertumbuhan video pendek Instagram Reels menjadi sumber informasi strategis, namun analisisnya belum optimal karena keterbatasan metode analisis. Penelitian ini bertujuan membandingkan metode pelabelan otomatis dan pengaruhnya terhadap klasifikasi *machine learning*. Data bersumber dari transkripsi 766 video Instagram Reels terkait program Makan Bergizi Gratis (MBG) menggunakan Whisper. Metodologi membandingkan dua pendekatan pelabelan, yaitu InSet dan DeepSeek, yang kemudian diklasifikasikan menggunakan algoritma Naive Bayes. Hasil pengujian skor *Cohen's kappa* dan *agreement* menunjukkan kualitas pelabelan DeepSeek lebih unggul daripada InSet. Melalui pengujian empat skenario pemodelan dan pertimbangan kualitas label, model *DeepSeek Imbalanced* terbukti paling optimal dengan capaian akurasi 66,7%, *macro F₁-score* 41,8%, serta karakteristik *good fit* pada *learning curve*.

Kata kunci: Analisis Sentimen, *Automatic Labeling*, Makan Bergizi Gratis (MBG), Instagram Reels, *Naive Bayes*.

ABSTRACT

PANDU PERSADA TANJUNG. A Comparison of Automatic Labelling in Naive Bayes Classification for Sentiment Analysis of Short Video Transcription Data. Supervised by YANI NURHADRYANI and HAFIDLOTUL FATIMAH AHMAD.

The growth of Instagram Reels short-form video content has become a strategic source of information. However, its analysis remains suboptimal due to limitations in existing analytical methods. This study aims to compare automatic labeling methods and their impact on machine learning classification performance. Data were derived from transcriptions of 766 Instagram Reels videos related to the Free Nutritious Meal (MBG) program, obtained using Whisper. The methodology compares two labeling approaches, namely InSet and DeepSeek, which were subsequently classified using the Naive Bayes algorithm. Testing results based on Cohen's kappa score and agreement level indicate that the labeling quality of DeepSeek is superior to that of InSet. Through the evaluation of four modeling scenarios and consideration of label quality, the DeepSeek Imbalanced model proved to be the most optimal, achieving an accuracy of 66.7%, a macro F1-score of 41.8%, and exhibiting good fit characteristics on the learning curve.

Keywords: Sentiment Analysis, Automatic Labelling, Free Nutritious Meals (MBG), Instagram Reels, Naive Bayes



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 20XX¹
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.

PERBANDINGAN *AUTOMATIC LABELLING* DALAM KLASIFIKASI *NAIVE BAYES* UNTUK ANALISIS SENTIMEN DATA TRANSKRIPSI VIDEO PENDEK

PANDU PERSADA TANJUNG

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

**PROGRAM STUDI ILMU KOMPUTER
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 Dr. Mushthofa, S.Kom, M.Sc



@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 *Automatic Labelling* dalam Klasifikasi Naive Bayes untuk Analisis Sentimen Data Transkripsi Video Pendek

Nama : Pandu Persada Tanjung

NIM : G6401221097

Disetujui oleh

Pembimbing 1:

Dr. Yani Nurhadryani S.Si, M.T

Pembimbing 2:

Hafidlotul Fatimah Ahmad, S.Kom., M.Kom.

Diketahui oleh

Ketua Program Studi Sarjana Ilmu Komputer:

Dr. Sony Hartono Wijaya, S.Kom., M.Kom.

19810809 200812 1 002

Tanggal Ujian:
8 Juli 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 Desember 2025 sampai bulan Mei 2026 ini ialah analisis sentimen, dengan judul “Perbandingan *Automatic Labelling* dalam Klasifikasi Naive Bayes untuk Analisis Sentimen Data Transkripsi Video Pendek”.

Penulis menyadari banyak pihak yang telah memberikan dukungan dan bantuan selama menyelesaikan tugas akhir ini. Pada kesempatan ini, penulis ingin menyampaikan terima kasih kepada:

- a. Orang tua penulis, Ibu Theresa Puti Tanjung yang selalu memberikan dukungan dan doa kepada penulis.
- b. Ibu Dr. Yani Nurhadryani S.Si, M.T dan Ibu Hafidlotul Fatimah Ahmad, S.Kom., M.Kom. selaku dosen pembimbing skripsi penulis yang senantiasa memberikan dukungan dan arahan selama penelitian.
- c. Bapak Herman, Bapak Yudha, Bapak Irvan, Bapak Ridwan selaku staf kependidikan yang telah membantu mempermudah segala urusan administrasi selama di IPB.
- d. Aryo Haris Wirakusuma Sudrajat selaku sahabat penulis yang selalu mendoakan dan mendukung proses pengerjaan tugas akhir ini.
- e. Ghiffari dan Darrel selaku rekan satu tim BDC Satria Data 2025 yang selalu mendukung proses pengerjaan tugas akhir satu sama lain.
- f. Adid, Naya, Salsa, Sindi, Wisnu, dan Yuda selaku teman penulis yang selalu memberikan bantuan dan dukungan disaat penulis mengalami kesulitan dalam pengerjaan tugas akhir ini.
- g. Fahmi, Candra, Rian, Bintang, Jo, dan Moses selaku teman asrama penulis yang selalu memberikan dukungan dalam pengerjaan tugas akhir ini.
- h. Annisa, Alya, Calvin, Dina, Farsa, Namira, dan Syarla selaku teman KKN penulis yang selalu memberikan dukungan dalam pengerjaan tugas akhir ini.
- i. Mahasiswa program studi S1 Ilmu Komputer angkatan 59 yang telah memberikan bantuan dan semangat kepada penulis selama perkuliahan.

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

Bogor, Juli 2026

Pandu Persada Tanjung

DAFTAR ISI

ABSTRAK	ii
DAFTAR TABEL	ix
DAFTAR GAMBAR	ix
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	2
1.3 Tujuan	2
1.4 Manfaat	3
1.5 Ruang Lingkup	3
II TINJAUAN PUSTAKA	4
2.1 Penelitian Sebelumnya	4
2.2 Landasan Teori	6
III METODE	14
3.1 Obyek Penelitian	14
3.2 Tahapan Penelitian	14
3.3 Lingkungan Pengembangan	18
IV HASIL DAN PEMBAHASAN	18
4.1 Persiapan Data Transkripsi Untuk Pelabelan dan Pelatihan Model	18
4.2 Pelatihan dan Evaluasi Model Klasifikasi Sentimen	31
4.3 Analisis Sentimen Data Transkripsi Dari Dua Metode Pelabelan	34
4.4 <i>Dashboard Untuk Scraping</i> , Klasifikasi, dan Visualisasi Sentimen	40
V SIMPULAN DAN SARAN	45
5.1 Simpulan	45
5.2 Saran	45
DAFTAR PUSTAKA	46
LAMPIRAN	50
RIWAYAT HIDUP	56



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 TABEL

1. Nilai dan kategori kappa	16
2. Perbandingan data untuk pelabelan leksikon	21
3. Perbandingan data untuk pelabelan LLM	22
4. Contoh data yang dilakukan validasi	28
5. Nilai <i>agreement</i> dan kappa pada metode pelabelan	29
6. Sebaran data pada tiap partisi sebelum SMOTE	29
7. Contoh hasil <i>feature extraction</i>	29

DAFTAR GAMBAR

1. <i>K-Fold Cross-Validation</i>	13
2. Diagram alur penelitian	15
3. Contoh data audio	19
4. <i>Wordcloud</i> dataset teks video	24
5. Diagram 15 unigram teratas	24
6. Diagram 15 bigram teratas	25
7. Diagram 15 trigram teratas	25
8. Hasil pelabelan leksikon	27
9. Hasil pelabelan LLM	27
10. Distribusi data sebelum SMOTE pada InSet Imbalanced	30
11. Distribusi data setelah SMOTE pada InSet SMOTE	30
12. Distribusi data sebelum SMOTE pada DeepSeek Imbalanced	31
13. Distribusi data setelah SMOTE pada DeepSeek SMOTE	31
14. Performa akurasi dan <i>macro f1-score</i> setiap model	32
15. <i>Learning curve</i> Inset Imbalanced	33
16. <i>Learning curve</i> Inset SMOTE	33
17. <i>Learning curve</i> Deepseek Imbalanced	34
18. <i>Learning curve</i> Deepseek SMOTE	34
19. Proporsi distribusi sentimen	34
20. Unigram pada pelabelan InSet dengan sentimen positif	35
21. Unigram pada pelabelan InSet dengan sentimen negatif	36
22. Bigram pada pelabelan InSet dengan sentimen positif	36
23. Bigram pada pelabelan InSet dengan sentimen negatif	36
24. Trigram pada pelabelan InSet dengan sentimen positif	37
25. Trigram pada pelabelan InSet dengan sentimen negatif	37
26. Unigram pada pelabelan DeepSeek dengan sentimen positif	38
27. Unigram pada pelabelan DeepSeek dengan sentimen negatif	38
28. Bigram pada pelabelan DeepSeek dengan sentimen positif	38
29. Bigram pada pelabelan DeepSeek dengan sentimen negatif	39
30. Trigram pada pelabelan DeepSeek dengan sentimen positif	39
31. Trigram pada pelabelan DeepSeek dengan sentimen negatif	39
32. Use case diagram aplikasi Shovia	40
33. Data flow diagram level 0	40
34. Data flow diagram level 1	41



35. Data flow diagram level 2	41
36. Data flow diagram level 2 dekomposisi 3.0	41
37. Halaman Beranda dashboard Shovia	42
38. Fitur scraping dan transkripsi dashboard Shovia	42
39. Hasil scraping dan fitur seleksi data dashboard Shovia	43
40. Fitur klasifikasi sentimen dashboard Shovia	43
41. Visualisasi proporsi kelas sentimen	43
42. Visualisasi dan analisis N-Gram	44
43. Tabel hasil klasifikasi sentimen	44