

PERFORMANCE EVALUATION OF INDOBERT AND LDA FOR SENTIMENT ANALYSIS AND TOPIC MODELING ON HALODOC USER REVIEWS

DHIYA KHALISHAH TSANY SUWARSO



**STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2025**



- 



STATEMENT REGARDING SKRIPSI, INFORMATION SOURCES, AND COPYRIGHT

Hereby I declare that the skripsi with the title “Performance Evaluation of IndoBERT and LDA for Sentiment Analysis and Topic Modeling on Halodoc User Reviews” is my work with guidance from the supervisory committee and it has not been submitted in any form to any university. Information sources that come or are quoted from published or unpublished work by other authors have been referred to in the text and mentioned in the bibliography. I hereby hand over the copyrights of my skripsi to IPB University.

Bogor, June 2025

Dhiya Khalishah Tsany Suwarso
G1401211038

- 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.



ABSTRAK

DHIYA KHALISHAH TSANY SUWARSO. Evaluasi Kinerja IndoBERT dan LDA untuk Analisis Sentimen dan Pemodelan Topik pada Ulasan Pengguna Aplikasi Halodoc. Dibimbing oleh KHAIRIL ANWAR NOTODIPUTRO dan SACHNAZ DESTA OKTARINA.

Pertumbuhan teknologi mendorong adopsi layanan kesehatan digital, seperti Halodoc yang memungkinkan pengguna melakukan konsultasi dokter dan pengiriman obat secara daring. Untuk memahami persepsi pengguna terhadap layanan ini, penelitian ini melakukan analisis sentimen menggunakan IndoBERT dan pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA) terhadap ulasan pengguna Halodoc di Google Play Store. Penelitian ini bertujuan mengevaluasi kinerja IndoBERT dalam mengklasifikasikan sentimen ulasan (positif, netral, negatif) serta mengidentifikasi topik-topik utama dalam ulasan positif dan negatif menggunakan LDA. Hasil penelitian menunjukkan bahwa IndoBERT dengan *class weight* memberikan performa terbaik dalam kondisi kelas tidak seimbang, dengan *macro F1-score* sebesar 0,8230 dan *balanced accuracy* sebesar 0,8274. Model ini mengklasifikasikan sentimen positif dan negatif dengan baik, tetapi performa pada sentimen netral rendah karena data terbatas, sifatnya ambigu, dan kosakata yang tumpang tindih. Selanjutnya, pemodelan topik dilakukan terhadap ulasan positif dan negatif. Model LDA dengan tokenisasi bigram menghasilkan nilai koherensi tertinggi dan topik yang lebih mudah diinterpretasikan karena mempertahankan pasangan kata yang bermakna. Pemilihan parameter K , α , dan β turut memengaruhi kualitas topik yang dihasilkan. Lima topik terbentuk dari ulasan positif dan tujuh topik dari ulasan negatif, dengan nilai koherensi masing-masing sebesar 0,6797 dan 0,6634. Temuan ini memberikan wawasan dalam memahami kebutuhan pengguna dan meningkatkan kualitas layanan Halodoc.

Kata kunci: analisis sentimen, halodoc, IndoBERT, *Latent Dirichlet Allocation* (LDA), pemodelan topik.

ABSTRACT

DHIYA KHALISHAH TSANY SUWARSO. Performance Evaluation of IndoBERT and LDA for Sentiment Analysis and Topic Modeling on Halodoc User Reviews. Supervised by KHAIRIL ANWAR NOTODIPUTRO and SACHNAZ DESTA OKTARINA.

Technological growth has driven the adoption of digital health services such as Halodoc, which allows users to consult doctors and receive medications online. To understand user perceptions of this service, this project conducts sentiment analysis using IndoBERT and topic modeling using Latent Dirichlet Allocation (LDA) on Halodoc user reviews from Google Play Store. This project aims to evaluate IndoBERT performance in classifying review sentiments (positive, neutral, negative) and to identify key topics in positive and negative reviews using LDA. Results show that IndoBERT with class weight achieved the best performance under class imbalance conditions, with macro F1-score of 0.8230 and balanced accuracy of 0.8274. The model effectively classified positive and negative sentiments, but struggled with neutral sentiment due to limited data, its ambiguity, and overlapping vocabulary. Subsequently, topic modeling was applied to the positive and negative reviews. LDA with bigram tokenization produced the highest coherence scores and more interpretable topics, by preserving meaningful word pairings. Parameter tuning of K , α , and β influenced topic quality. Five topics emerged from positive reviews and seven from negative reviews, with coherence scores of 0.6797 and 0.6634, respectively. These findings offer valuable insights into user needs and can support service improvement for Halodoc.

Keywords: halodoc, IndoBERT, Latent Dirichlet Allocation (LDA), sentiment analysis, topic modeling.



- 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.

© Copyright IPB University, 2025
Copyrights protected by Law

Prohibited to cite parts or all of this manuscript without referencing its source. Citations are only permitted for educational purposes, research, writing of scientific manuscripts, reports, critiques, or reviews. Citations may not harm IPB University's interests.

Publishing and copying some or parts of this manuscript without authorization from IPB University is prohibited.

PERFORMANCE EVALUATION OF INDOBERT AND LDA FOR SENTIMENT ANALYSIS AND TOPIC MODELING ON HALODOC USER REVIEWS

DHIYA KHALISHAH TSANY SUWARSO

Skripsi
to complete the requirement for graduation of
Bachelor Degree in
Statistics and Data Science Study Program

**STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
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.

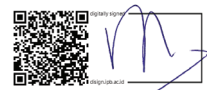


@Hak cipta milik IPB University

Title : Performance Evaluation of IndoBERT and LDA for Sentiment Analysis and Topic Modeling on Halodoc User Reviews
Name : Dhiya Khalishah Tsany Suwarso
Student ID : G1401211038

Approved by

Main Supervisor:
Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S.



Co-Supervisor:
Sachnaz Desta Oktarina, S.Stat., M.Agr.Sc., Ph.D.



Acknowledged by

Head of Statistics and Data Science Study Program:
Dr. Bagus Sartono, S.Si., M.Si
NIP 19780411 200501 1002



Date of Thesis Defense:
June 3rd 2025

Date of Graduation:

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 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.

PREFACE

Praise and gratitude are sincerely expressed to Allah subhanaahu wa ta'ala for all His blessings, which have enabled the completion of this scientific work. The theme chosen for the project conducted from January 2025 to June 2025 is sentiment analysis and topic modeling, titled "Performance Evaluation of IndoBERT and LDA for Sentiment Analysis and Topic Modeling on Halodoc User Reviews".

The completion of this scientific work is due to various support and assistance from different parties. Therefore, the author would like to thank all parties who have helped in the completion of this project, including:

1. The author's parents (Gatot Suwarso and Popy Puspitasari), siblings, and extend family for their boundless support and prayers.
2. The author's advisors, Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S. and Sachnaz Desta Oktarina, S.Stat, M.Agr.Sc., Ph.D. who have assisted, guided, directed, and supported during the preparation of this project.
3. The external examiner, Aulia Rizki Firdawanti, S.Stat., M.Si., who provided critical evaluation and thoughtful recommendations, which have substantially contributed to enhancing the clarity of this project.
4. The moderators, Akbar Rizki, S.Stat., M.Si., and Gerry Alfa Dito, S.Si., M.Si., who have offered constructive feedback and insightful suggestions during the colloquium and seminar, contributing to the improvement of this project.
5. All the lecturers and staff of the Department of Statistics who have imparted valuable knowledge and supported all the author's needs during the coursework and the preparation of this scientific work.
6. All friends, especially Angel, Diva, Kheni, Reyza, Fuadiyah, Kamilah, Adiba, Arfiah, Oktavia, Azra and all peers from Statistics Batch 58, for their invaluable support and solidarity throughout this academic journey.
7. All parties who have assisted the author in preparing this scientific work, who cannot be mentioned one by one.

The author acknowledges that this scientific work is still not perfect. Therefore, the author apologizes if there are still any mistakes or weaknesses. May this scientific work benefit those who need it and contribute to the advancement of knowledge.

Bogor, June 2025

Dhiya Khalishah Tsany Suwarso



@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.

TABLE OF CONTENTS

LIST OF TABLES	xii
LIST OF FIGURES	xii
LIST OF APPENDICES	xiii
I INTRODUCTION	1
1.1 Background	1
1.2 Objectives	2
II LITERATURE REVIEW	3
2.1 Sentiment Analysis	3
2.2 IndoBERT	3
2.3 Synthetic Minority Over-sampling Technique (SMOTE)	6
2.4 Class Weight	6
2.5 Hyperparameter Tuning with Grid Search	7
2.6 Stratified k -Fold Cross Validation	7
2.7 Evaluation of Sentiment Analysis Model	8
2.8 Topic Modeling	9
2.9 Latent Dirichlet Allocation (LDA)	10
2.10 Variational Inference	11
2.11 Tokenization N-gram	12
2.12 Bag-of-Words	13
2.13 Topic Coherence	13
III METHOD	15
3.1 Data	15
3.2 Data Analysis Procedure	15
IV RESULTS AND DISCUSSION	21
4.1 Data Preprocessing	21
4.2 Data Labeling	22
4.3 Data Exploration	24
4.4 Sentiment Analysis Using IndoBERT	25
4.4.1 Baseline Sentiment Classification Using IndoBERT	25
4.4.2 Sentiment Classification Using IndoBERT with SMOTE	28
4.4.3 Sentiment Classification Using IndoBERT with Class Weight	30
4.4.4 Comparison of Sentiment Classification Models	32
4.5 Topic Modeling Using Latent Dirichlet Allocation (LDA)	33
4.5.1 N-gram Tokenization	33
4.5.2 Bag-of-Words (BoW)	34
4.5.3 Latent Dirichlet Allocation on Positive Sentiment Reviews	35
4.6 Topic Interpretation by Sentiment Category	40
4.6.1 Topic Interpretation – Positive Sentiment	40
4.6.2 Topic Interpretation – Negative Sentiment	43
4.7 Analysis of Response Time Differences Based on Review Sentiment	46
V CONCLUSION AND RECOMMENDATION	48
5.1 Conclusion	48
5.2 Recommendation	48
REFERENCES	50

APPENDICES
BIOGRAPHY

@Hak cipta milik IPB University

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.

LIST OF TABLES		
1	Architecture spesification of BERT _{BASE} and BERT _{LARGE} Models	4
2	Multi-class classification confusion matrix (Markoulidakis <i>et al.</i> 2021)	8
3	Sample of user review data collected	15
4	Hyperparameter values	19
5	Sample of preprocessed text data	21
6	Sample of review sentiment labels	22
7	Distribution of positive, neutral, and negative labels by method	23
8	Hyperparameter tuning results for the IndoBERT model	26
9	Classification report of the IndoBERT model	26
10	Hyperparameter tuning results for the IndoBERT SMOTE model	28
11	Classification report of the IndoBERT SMOTE model	29
12	Hyperparameter tuning results for the IndoBERT class weight model	31
13	Classification report of the IndoBERT class weight model	31
14	Performance Comparison of Sentiment Classification Models	32
15	Illustration of N-gram tokenization on preprocessed review	33
16	Bag-of-Words and top terms in positive reviews	34
17	Bag-of-Words and top terms in negative reviews	35
18	Coherence scores for positive reviews topic modeling	36
19	Coherence scores for negative reviews topic modeling	39
20	LDA topic modeling results on positive sentiment reviews	41
21	LDA topic modeling results on negative sentiment reviews	43

LIST OF FIGURES		
1	Illustration of pre-training and fine-tuning in BERT (Devlin <i>et al.</i> 2019)	4
2	BERT input representation (Devlin <i>et al.</i> 2019)	5
3	Illustration of the SMOTE algorithm (Graa and Rekik 2019)	6
4	Illustration of the 5-fold cross-validation (Setiawan <i>et al.</i> 2023)	8
5	Graphical model representation of LDA (Blei <i>et al.</i> 2003)	10
6	Project Flowchart of the Sentiment Analysis and Topic Modeling	16
7	Sentiment class distribution in user reviews	23
8	Sentiment wordclouds (a) Positive (b) Neutral (c) Negative	24
9	Confusion matrix of the IndoBERT model on the test set	27
10	Class distribution before and after SMOTE for each fold	28
11	Confusion matrix of the IndoBERT SMOTE model on the test set	30
12	Confusion matrix of the IndoBERT class weight model on the test set	32
13	Coherence score trends of LDA bigram for positive reviews	37
14	LDA Bigram topic visualization for positive reviews	37
15	Coherence score trends of LDA bigram for negative reviews	39
16	LDA Bigram topic visualization for negative reviews	40
17	Boxplot of response time based on sentiment	46



LIST OF APPENDICES

1	Word cloud visualization of LDA topics in positive sentiment	55
2	Word cloud visualization of LDA topics in negative sentiment	55

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.