

TRANSFORMER-BASED MODEL OPTIMIZATION OF INDOBERT AND XLNET FOR INDONESIAN HOAX NEWS DETECTION

FARREL GILBRAN



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





STATEMENT REGARDING UNDERGRADUATE THESIS, INFORMATION SOURCES, AND COPYRIGHT

Hereby I declare that the skripsi with the title “Transformer-Based Model Optimization of IndoBERT and XLNet for Indonesian Hoax News Detection” 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 Information sources that come or are quoted from published or unpublished work by other authors have been referred to in the text and mentioned in the bibliography.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, August 2025

Farrel Gilbran
G1401211057

- 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

FARREL GILBRAN. Optimalisasi Model Berbasis Transformer IndoBERT dan XLNet untuk Deteksi Berita Hoaks Berbahasa Indonesia. Dibimbing oleh ASEP SAEFUDDIN dan SEPTIAN RAHARDIANTORO.

Penyebaran berita hoaks di internet telah menjadi ancaman serius yang dapat memengaruhi opini publik, menciptakan ketegangan sosial, dan merusak reputasi individu maupun kelompok. Masyarakat Anti-Fitnah Indonesia (MAFINDO) telah melakukan identifikasi berita hoaks secara manual melalui platform turnbackhoax.id. Namun, metode ini menjadi tidak efisien seiring dengan meningkatnya volume informasi. Oleh karena itu, dibutuhkan solusi berbasis machine learning untuk mengotomatisasi dan meningkatkan deteksi berita hoaks. Penelitian ini membandingkan performa model berbasis transformer, yaitu IndoBERT dan XLNet, dalam mengklasifikasikan berita hoaks berbahasa Indonesia. Dataset yang digunakan terdiri atas 18689 artikel hoaks dari turnbackhoax.id dan 10229 artikel non-hoaks yang diambil dari detik.com. Proses analisis melibatkan tahapan praproses seperti *filtering*, *case folding*, dan tokenisasi sebelum data digunakan untuk pemodelan. Untuk mengatasi potensi ketidakseimbangan data dan memastikan evaluasi yang andal, digunakan *stratified 3-fold cross-validation*. Optimasi model dilakukan dengan menggunakan Optuna, sebuah *framework* efisien untuk optimasi *hyperparameter*. Hasil penelitian menunjukkan bahwa model transformer berbahasa Indonesia, khususnya IndoBERT, memiliki performa yang sangat baik dalam mendeteksi berita hoaks. IndoBERT mencapai performa terbaik dengan akurasi sebesar 0.9924, presisi 0.9968, recall 0.9914, dan skor F1 sebesar 0.9941. Meskipun XLNet juga menunjukkan performa yang kuat, model ini sedikit kurang efektif dibandingkan IndoBERT, yang menunjukkan bahwa XLNet mungkin kurang optimal untuk tugas berbahasa Indonesia tanpa adaptasi khusus. Secara keseluruhan, penelitian ini menunjukkan bahwa *stratified cross-validation* dan *tuning hyperparameter* yang tepat dapat secara efektif mengatasi ketidakseimbangan data. Selain itu, perbedaan panjang token antara kelas hoaks dan non-hoaks memengaruhi proses pembelajaran model, yang menegaskan pentingnya mempertimbangkan baik jumlah sampel maupun kompleksitas teks.

Kata kunci: berita, hoaks, indobert, optuna, xlnet

ABSTRACT

FARREL GILBRAN. Transformer-Based Model Optimization of IndoBERT and XLNet for Indonesian Hoax News Detection. Supervised by ASEP SAEFUDDIN and SEPTIAN RAHARDIANTORO.

The spread of hoax news on the internet has become a serious threat that can influence public opinion, create social tensions, and damage the reputation of individuals or groups. Masyarakat Anti-Fitnah Indonesia (MAFINDO) has manually identified hoax news through the platform turnbackhoax.id. However, this method is inefficient as the volume of information increases. Therefore, a machine learning-based solution is needed to automate and improve hoax news detection. This study compares the performance of transformer-based models, namely IndoBERT and XLNet, in classifying Indonesian-language hoax news. The dataset consists of 18689 hoax articles from turnbackhoax.id and 10229 non-hoax articles scraped from Detik.com. The process involves preprocessing steps such as filtering, case folding, and tokenization before the data is used for modeling. To address potential data imbalance and ensure robust evaluation, stratified 3-fold cross-validation was employed. Model optimization was carried out using Optuna, an efficient hyperparameter optimization framework. The results show that Indonesian-language transformer models, particularly IndoBERT, perform exceptionally well in hoax news detection. IndoBERT achieved the best performance with an accuracy of 0.9924, precision of 0.9968, recall of 0.9914, and F1 score of 0.9941. Although XLNet also demonstrated strong performance, it was slightly less effective than IndoBERT, suggesting it may be less optimal for Indonesian-language tasks without specific adaptation. Overall, the study shows that stratified cross-validation and proper hyperparameter tuning can effectively mitigate data imbalance. Moreover, differences in token length between hoax and non-hoax classes influence model learning, highlighting the importance of accounting for both sample size and textual complexity.

Keywords: hoax, indobert, news, optuna, xlnet



- 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.



TRANSFORMER-BASED MODEL OPTIMIZATION OF INDOBERT AND XLNET FOR INDONESIAN HOAX NEWS DETECTION

FARREL GILBRAN

Undergraduate Thesis
To complete the requirement for graduation of
Bachelor Degree in
Statistics and Data Science Study Program

**STATISTICS AND DATA SCIENCE BACHELOR 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.

Title : Transformer-Based Model Optimization of IndoBERT and XLNet
for Indonesian Hoax News Detection
Name : Farrel Gilbran
Student ID : G1401211057

Approved by

Main Supervisor:
Prof. Dr. Ir. Asep Saefuddin, M.Sc.



Co-Supervisor:
Dr. Septian Rahardiantoro S.Stat., M.Si.



Acknowledged by

Head of Statistics and Data Science Study Program:
Dr. Bagus Sartono, S.Si., M.Si.
NIP 197804112005011002



Date of Thesis Defense:
June 30, 2025

Date of Graduation:



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 February 2025 to June 2025 is modelling hoax news detection, titled "Transformer-Based Model Optimization of IndoBERT and XLNet for Indonesian Hoax News Detection".

The author would like to express sincere thanks to all parties who have offered prayers, guidance, and support during the preparation of this scientific work, including:

1. Mr. Arwen and Mrs. Susi as my beloved parents, as well as Diego, Angel, Leo, and Sophie as my siblings; Mr. Aan and Mrs. Yelvi as my uncle and aunt; and Aurel, Fahri, and Maulana as my cousins, who have given their prayers, encouragement, love, and full support to the author.
2. Prof. Dr. Ir. Asep Saefuddin, M.Sc., and Dr. Septian Rahardianto, S.Stat., M.Si., as the academic supervisors who continuously provided advice and guidance to the author in completing this scientific work.
3. Prof. Dr. Ir. Aji Hamim Wigena, M.Sc as the external examiner, and also Mrs. Akbar Rizki S.Stat., M.Si. and Dr. Anwar Fitrianto S.Si., M.Sc. as the moderators of the colloquium and seminar, who have provided valuable feedback and suggestions for improving this work.
4. All lecturers and the academic community of the Statistics and Data Science Study Program, SSMI IPB, who have shared useful knowledge and lessons with the author, as well as the administrative staff who supported the author in handling academic processes.
5. Much, Vita, Okta, Farhan, Fajar, and Farras as members of the Rumah Guru Perdesi residence and close friends of the author who have always been there since the PPKU period, offering prayers and continuous support throughout the journey.
6. Zahra, Ignés, Dewi, and Radhit known as the "musuhignes", who are the author's close friends that have always been present throughout the university years, offering prayers and unwavering support.
7. Asathin, Viona, and Anggun, the author's high school friends who have always given support in many ways despite the distance.
8. All members of Statistics 58 especially Uci, Ara, Fedo and everyone who has contributed and shown kindness to the author, although not mentioned one by one, whose support helped bring this scientific work to completion.

May this scientific work be useful for those in need and contribute to the advancement of knowledge.

Bogor, August 2025

Farrel Gilbran



@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.

CONTENTS

LIST OF TABLES	xi
LIST OF FIGURES	xi
LIST OF APPENDICES	xi
I INTRODUCTION	1
1.1 Background	1
1.2 Objectives	2
II LITERATUR REVIEW	4
2.1 Hoax and Non-Hoax News	4
2.2 Web Crawling and Web Scrapping	4
2.3 Natural Language Processing	5
2.4 IndoBERT	5
2.5 XLNet	8
2.6 Optuna	12
2.7 Stratified K-Fold Cross Validation	13
III METHODOLOGY	15
3.1 Data	15
3.2 Data Analysis Procedure	16
IV RESULT AND DISCUSSION	19
4.1 Data Preprocessing	19
4.2 Data Eksploration	20
4.3 Training and Best Parameter Search of IndoBERT and XLNet	21
4.4 Final Training and Evaluation of IndoBERT and XLNet	22
4.5 Implication	24
V CONCLUSION AND RECOMMENDATION	26
5.1 Conclusions	26
5.2 Recommendation	26
REFERENCES	27
APPENDICES	30
BIOGRAPHY	31

LIST OF TABLES

1	Table 1 Example of Crawled Data	15
2	Table 2 Confusion matrix	17
3	Table 3 Preprocessing Result	19
4	Table 4 Statistics Descriptive of Label Tokens	20
5	Table 5 Best Parameter Model	21
6	Table 6 Final Model Train Evaluation	22
7	Table 7 Final Model Test Evaluation	22

LIST OF FIGURES

1	Figure 1 Illustration of pre-training and fine-tuning in BERT (Devlin et al. 2019)	7
2	Figure 2 Overview of the permutation language modeling training with two-stream attention	11
3	Figure 3 Schematic diagram Stratified K-fold Cross Validation	14
4	Figure 4 Wordclouds for Label Non-Hoax (0) and Label Hoax (1)	20
5	Figure 5 IndoBERT Loss & F1-Score Plots for Training and Validation	23
6	Figure 6 XLNet Loss & F1-Score Plots for Training and Validation	23
7	Figure 7 Confusion matrix of the IndoBERT(a) and XLNet(b) model on the test set	24

LIST OF APPENDICES

1	Appendices 1 IndoBERT and XLNet Parameter Trials Tuning	30
2	Appendices 2 IndoBERT and XLNet Evaluation fo Each Trials Tuning	30