PENGKLASIFIKASIAN TEKS DARI MEDIA SOSIAL X TERHADAP DEBAT CAPRES INDONESIA TAHUN 2024 MENGGUNAKAN METODE IndoBERT

SKRIPSI

untuk memenuhi persyaratan
dalam menyelesaikan program sarjana Strata-1 Statistika

Oleh
MUHAMMAD NABRIS SABANA
NIM. 2111017310007

PROGRAM STUDI S-1 STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JULI 2025

SKRIPSI

PENGKLASIFIKASIAN TEKS DARI MEDIA SOSIAL *X* TERHADAP DEBAT CAPRES INDONESIA TAHUN 2024 MENGGUNAKAN METODE IndoBERT

Oleh
Muhammad Nabris Sabana
NIM. 2111017310007

Telah dipertahankan pada hari Kamis, tanggal 10 Juli 2025 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

Pembimbing I

Yuana Sukmawaty, S.Si., M.Si
NIP. 198810152015042002

Penguji I

Dewi Sri Susanti, S.Si., M.Si
NIP. 197305161999032002

Pembimbing II

Selvi Annisa, S.Si., M.Si
NIP. 199212262022032016

Penguji II

Maisarah, S.Pd.I, M.Pd
NIP. 19890713201801213001

Banjarbaru, 21 Juli 2025
Mengetahui,
Koordinator Program Studi Statistika FMIPA ULM

Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D
NIP. 198303282005012001

SKRIPSI

PENGKLASIFIKASIAN TEKS DARI MEDIA SOSIAL *X* TERHADAP DEBAT CAPRES INDONESIA TAHUN 2024 MENGGUNAKAN METODE IndoBERT

Oleh
Muhammad Nabris Sabana
NIM. 2111017310007

Telah dipertahankan pada hari Kamis, tanggal 10 Juli 2025 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

Pembimbing I	**Penguji I**
Yuana Sukmawaty, S.Si., M.Si	*Dewi Sri Susanti, S.Si., M.Si*
NIP. 198810152015042002	*NIP. 197305161999032002*
Pembimbing II	**Penguji II**
Selvi Annisa, S.Si., M.Si	*Maisarah, S.Pd.I, M.Pd*
NIP. 199212262022032016	*NIP. 198907132018012130001*

Banjarbaru, 21 Juli 2025
Mengetahui,
Koordinator Program Studi Statistika FMIPA ULM

Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D
NIP. 198303282005012001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 21 Juli 2025

Muhammad Nabris Sabana

NIM. 2111017310007

ABSTRAK

Pengklasifikasian Teks Dari Media Sosial X Terhadap Debat Capres Indonesia Tahun 2024 Menggunakan Metode IndoBERT *(Oleh : Muhammad Nabris Sabana; Pembimbing: Yuana Sukmawaty dan Selvi Annisa, 2025; 52 Halaman)*

Media sosial X menjadi salah satu platform diskusi terpopuler bagi masyarakat utama dalam memberikan opini dan respons terhadap isu-isu politik seperti debat calon presiden Indonesia tahun 2024. Para capres saling beradu argumen dalam debat sehingga masyarakat berdiskusi di media sosial X untuk memberikan respons terhadap salah satu capres. Selain itu, sering kali tweet diskusi berisikan respons dengan beberapa aspek kenegaraan seperti ekonomi, kebudayaan, politik, dan lain-lainnya. Karena banyaknya aspek kenegaraan, sehingga harus dikategorikan dengan bantuan analisis. Permasalahan ini bisa diatasi dengan menggunakan metode klasifikasi teks IndoBERT. Penelitian ini bertujuan untuk mengimplementasikan metode IndoBERT untuk mengklasifikasikan teks yang diambil dari media sosial X terhadap debat calon presiden Indonesia Tahun 2024. Pembentukan model diawali dengan pembuatan visualisasi wordcloud untuk mengetahui topik respons masyarakat yang sering dibicarakan, kemudian pembagian data latih dan uji. Data latih kemudian diterapkan teknik SMOTE-N untuk menyeimbangkan antar kelasnya, setelah itu dilakukan fine-tuning dengan beberapa kombinasi parameter. Hasil dari pelatihan akan dievaluasi menggunakan data uji. Dari hasil evaluasi, diperoleh nilai balanced accuracy sebesar 0.474452 dengan tingkat klasifikasi cukup.

Kata kunci: Media Sosial X, IndoBERT, Debat Capres, SMOTE-N, Wordcloud

ABSTRACT

Classifying Texts from Social Media X on Indonesia's 2024 Presidential Candidate Debate Using the IndoBERT Method *(By : Muhammad Nabris Sabana; Supervisors: Yuana Sukmawaty and Selvi Annisa, 2025; 52 Pages)*

Social media X has become one of the most popular discussion platforms for mainstream society to provide opinions and responses to political issues such as the Indonesian presidential candidate debate in 2024. The presidential candidates argued with each other in the debate, so people discussed on social media X to respond to one of the presidential candidates. In addition, discussion tweets often contain responses with several aspects of the state such as economy, culture, politics, and others. Because there are many aspects of the state, so it must be categorized with the help of analysis. This problem can be solved by using the IndoBERT text classification method. This research aims to implement the IndoBERT method to classify text taken from X social media on the Indonesian presidential candidate debate in 2024. Model building begins with the creation of wordcloud visualization to find out the topic of public response that is often discussed, then the division of training and test data. The training data is then applied the SMOTE-N technique to balance between classes, after which fine-tuning is carried out with several combinations of parameters. The results of the training will be evaluated using test data. From the evaluation results, a balanced accuracy value of 0.474452 is obtained with a sufficient classification level.

Keyword: Social Media X, IndoBERT, Presidential Candidate Debate, SMOTE-N, Wordcloud

PRAKATA

Alhamdulillahi robbil 'alamin, puji dan syukur penulis ucapkan ke hadirat Allah SWT yang telah memberikan rahmat, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "Pengklasifikasian Teks Dari Media Sosial X Terhadap Debat Capres Indonesia Tahun 2024 Menggunakan Metode IndoBERT". Tugas Akhir ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan program sarjana strata-1 di Program Studi Statistika FMIPA ULM.

Dalam penyusunan Tugas Akhir ini, penulis mendapatkan berbagai dukungan serta bantuan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. *Kedua orang tua dan keluarga yang selalu memberikan doa, dukungan materil, nasihat, dan motivasi dalam proses penyusunan Tugas Akhir.*
2. *Ibu Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D selaku Koordinator Program Studi Statistika FMIPA ULM atas segala ilmu dan dukungan yang telah diberikan selama perkuliahan di Program Studi Statistika FMIPA ULM.*
3. *Ibu Yuana Sukmawaty, S.Si., M.Si dan Ibu Selvi Annisa, S.Si., M.Si selaku dosen pembimbing yang telah memberikan bimbingan dan arahan serta dukungan dalam proses penyusunan Tugas Akhir.*
4. *Ibu Dewi Sri Susanti, S.Si, M.Si dan Ibu Maisarah, S.Pd.I, M.Pd selaku dosen penguji yang telah memberikan masukan, saran, serta bantuan dalam proses perbaikan Tugas Akhir.*
5. *Seluruh dosen dan staf Program Studi Statistika FMIPA ULM yang telah memberikan ilmu, nasihat, serta dukungan selama masa perkuliahan di Program Studi Statistika FMIPA ULM.*
6. *Teman-teman yang selalu menemani, mendukung, serta mendengarkan keluh-kesah saya selama penulisan Tugas Akhir maupun selama masa perkuliahan, khususnya teman-teman "El Juara" dan "Serangan Fajar".*
7. *Teman-teman Program Studi Statistika FMIPA ULM khususnya Angkatan 2021 yang telah berjuang bersama selama perkuliahan.*
8. *Last but not least, I wanna thank me.*

Penulisan Tugas Akhir ini tentu masih terdapat kekurangan dan jauh dari kata sempurna, sehingga masukan, kritik, dan saran yang membangun dari semua pihak sangat diharapkan dapat membantu dalam penyempurnaan Tugas Akhir ini. Akhir kata, penulis berharap agar hasil penelitian pada Tugas Akhir ini dapat bermanfaat bagi semua pihak.

Banjarbaru, 21 Juli 2025

Muhammad Nabris Sabana

DAFTAR ISI

PERNYATAAN ..... *iv*
ABSTRAK ..... *v*
ABSTRACT ..... *vi*
PRAKATA ..... *vii*
DAFTAR ISI ..... *viii*
DAFTAR GAMBAR ..... *x*
DAFTAR TABEL ..... *xi*
DAFTAR LAMPIRAN ..... *xii*
DAFTAR ISTILAH ..... *xiii*
BAB I PENDAHULUAN *1*
1.1 Latar Belakang *1*
1.2 Rumusan Masalah *3*
1.3 Tujuan Penelitian *3*
1.4 Manfaat Penelitian *3*
BAB II TINJAUAN PUSTAKA *4*
2.1 Penelitian Terdahulu *4*
2.2 Kajian Teori *8*
2.2.1 Klasifikasi Teks *8*
2.2.2 X (Twitter) *8*
2.2.3 Natural Languange Processing *9*
2.2.4 Text Pre-processing *9*
2.2.5 Wordcloud *11*
2.2.6 Transformers *11*
2.2.7 Bidirectional Encoders Representations from Transformers *16*
2.2.8 Penanganan Imbalance Data *21*
2.2.9 Multiclass Classification Confusion Matrix *22*
2.2.10 Balanced Accuracy *22*
BAB III METODE PENELITIAN *24*
3.1 Sumber Data *24*
3.2 Variabel Penelitian *24*
3.3 Prosedur Penelitian *24*
BAB IV HASIL DAN PEMBAHASAN *28*
4.1 Persiapan Data dan Visualisasi Deskriptif *28*
4.1.1 Text Pre-processing *28*
4.1.2 Wordcloud *31*
4.2 Penerapan Model IndoBERT *39*
4.2.1 Tokenisasi Teks dengan Pretrained IndoBERT *41*
4.2.2 Input Formatting *42*
4.2.3 Fine-tuning Model IndoBERT *44*
4.3 Evaluasi Hasil Klasifikasi *46*
BAB V PENUTUP *51*
5.1 Kesimpulan *51*
5.2 Saran *52*
DAFTAR PUSTAKA *53*

LAMPIRAN..... *57*
RIWAYAT HIDUP..... *92*

DAFTAR GAMBAR

Gambar 2.1 Arsitektur Transformers.....12
Gambar 2.2 Arsitektur Scaled Dot-Product Attention.....13
Gambar 2.3 Arsitektur Multi-Head Attention.....14
Gambar 2.4 Arsitektur Bidirectional Self-attention.....17
Gambar 2.5 Arsitektur Masked Languange Model.....18
Gambar 2.6 Ilustrasi Pre-training dan Fine-tuning.....20
Gambar 3.1 Diagram Prosedur Penelitian.....27
Gambar 4.1 Wordcloud Label Demografi.....31
Gambar 4.2 Wordcloud Label Ekonomi.....32
Gambar 4.3 Wordcloud Label Geografi.....33
Gambar 4.4 Wordcloud Label Ideologi.....34
Gambar 4.5 Wordcloud Label Pertahanan & Keamanan.....35
Gambar 4.6 Wordcloud Label Politik.....36
Gambar 4.7 Wordcloud Label Sosial Budaya.....37
Gambar 4.8 Wordcloud Label Sumber Daya Alam.....38
Gambar 4.9 Proporsi Label Klasifikasi Data Keseluruhan.....39
Gambar 4.10 Proporsi Label Klasifikasi Data Latih.....40
Gambar 4.11 Proporsi Label Klasifikasi Data Uji.....41
Gambar 4.12 (a) Matriks Input IDs; (b) Token Type IDs; (c) Attention Mask.....44

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu.....4
Tabel 2.2 Tabel Multiclass Classification Confusion Matrix.....22
Tabel 3.1 Variabel Penelitian.....24

Tabel 4.1 Hasil Penerapan Casefolding 29
Tabel 4.2 Hasil Penerapan Filtering 29
Tabel 4.3 Hasil Penerapan Tokenizing 30
Tabel 4.4 Hasil dari Data Cleaning 31
Tabel 4.5 Jumlah Label Data Asli 39
Tabel 4.6 Jumlah Label Klasifikasi Data Latih 40
Tabel 4.7 Jumlah Per Label Data Uji 40
Tabel 4.8 Jumlah Label Klasifikasi Setelah Penerapan SMOTE-N 41
Tabel 4.9 Tokenisasi Teks Pretrained IndoBERT 42
Tabel 4.10 Parameter Fine-tuning 44
Tabel 4.11 Balanced Accuracy Batch Size 16 45
Tabel 4.12 Confusion Matrix Parameter Terbaik 46
Tabel 4.13 Ringkasan True Positive Setiap Label 47
Tabel 4.14 Ringkasan False Negative Setiap Label 47
Tabel 4.15 Ringkasan Recall Setiap Label 48
Tabel 4.16 Perbandingan BA Tanpa SMOTE-N dengan SMOTE-N 49
Tabel 4.17 Perbandingan Label Pada Aktual dan Prediksi 49
Tabel 4.18 Rangkuman Kecenderungan Prediksi Label 49

DAFTAR LAMPIRAN

Lampiran 1. Data Hasil Scrapping dari Media Sosial X 57
Lampiran 2. Data Setelah di Text Pre-processing 60
Lampiran 3. Kata Kunci Stopword Umum/Kata Ganti 62

Lampiran 4. Kata Kunci Stopword Salah Tulis.....63
Lampiran 5. Kata Kunci Stopword Nama Orang, Lokasi, dan Nama Tempat....65
Lampiran 6. Hasil Klasifikasi Semua Kombinasi Parameter.....66
Lampiran 7. Perhitungan False Negative.....69
Lampiran 8. Perhitungan Recall Setiap Label.....70
Lampiran 9. Syntax Python untuk Install Library dan Input Data.....71
Lampiran 10. Syntax Python Text Preprocessing Casefolding dan Filtering.....72
Lampiran 11. Syntax Python Text Preprocessing Data Cleaning.....73
Lampiran 12. Syntax Python Text Preprocessing Tokenizing.....74
Lampiran 13. Syntax Python Text Preprocessing Stopword.....75
Lampiran 14. Syntax Python Text Preprocessing Stemming.....78
Lampiran 15. Syntax Python Wordcloud.....80
Lampiran 16. Syntax Python Pembagian Data.....81
Lampiran 17. Syntax Python Visualisasi Data untuk Melihat Proporsi Label....82
Lampiran 18. Syntax Python Penerapan Teknik SMOTE-N.....83
Lampiran 19. Syntax Python Input Formatting.....85
Lampiran 20. Syntax Python Visualisasi Matriks.....87
Lampiran 21. Syntax Python Fine-Tuning IndoBERT.....89
Lampiran 22. Syntax Python Mengevaluasi Hasil Klasifikasi.....91

DAFTAR LAMBANG DAN SINGKATAN

Q	*: Matriks koleksi vektor pertanyaan (queries)*
K	**:** *Matriks koleksi vektor kunci (keys)*

V	**:** *Matriks koleksi vektor nilai (values)*
pos	**:** *Posisi token dalam urutan*
d_{model}	*: Dimensi dari positional encoding*
x (FFN)	*: Vektor input yang memiliki dimensi* d_{model}
W_1	*: Matriks bobot pertama yang digunakan untuk transformasi linier pertama*
b_1	*: Vektor bias yang ditambahkan setelah transformasi linier pertama*
F_1, F_0	*: Dua nilai fitur yang sesuai*
C_1	*: Total jumlah kemunculan dari nilai fitur* F_1
C_{1i}	*: Jumlah kemunculan dari nilai fitur* F_1 *untuk kelas i*
C_0	*: Total jumlah kemunculan dari nilai fitur* F_0
C_{0i}	*: Jumlah kemunculan dari nilai fitur* F_0 *untuk kelas i*
k	*: Merupakan nilai konstanta yang biasanya bernilai 1*
TP	*: Jumlah yang benar di mana kelas positif terprediksi dengan benar*
FN	*: Jumlah yang salah di mana kelas positif diprediksi sebagai negatif*