

**ANALISIS SENTIMEN KOMENTAR YOUTUBE
BERBAHASA INDONESIA TERKAIT ISU WARGA NEGARA
INDONESIA DI JEPANG MENGGUNAKAN MODEL
HYBRID INDOBERT DAN LSTM**



SKRIPSI

**VANESS FEBRAIO
00000055310**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

**ANALISIS SENTIMEN KOMENTAR YOUTUBE
BERBAHASA INDONESIA TERKAIT ISU WARGA NEGARA
INDONESIA DI JEPANG MENGGUNAKAN MODEL
HYBRID INDOBERT DAN LSTM**



Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

**VANESS FEBRAIO
00000055310**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Vaness Febraio
Nomor Induk Mahasiswa : 00000055310
Program Studi : Informatika

Skripsi dengan judul:

Analisis Sentimen Komentar YouTube Berbahasa Indonesia terkait Isu Warga Negara Indonesia di Jepang Menggunakan Model Hybrid IndoBERT dan LSTM

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 22 Desember 2025



(Vaness Febraio)

HALAMAN PENGESAHAN

Skripsi dengan judul

ANALISIS SENTIMEN KOMENTAR YOUTUBE BERBAHASA INDONESIA TERKAIT ISU WARGA NEGARA INDONESIA DI JEPANG MENGUNAKAN MODEL HYBRID INDOBERT DAN LSTM

oleh

Nama : Vaness Febraio
NIM : 00000055310
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Kamis, 15 Januari 2026
Pukul 08.00 s/s 10.00 dan dinyatakan
LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang



(David Agustriawan, S.Kom., M.Sc.,
Ph.D.)

NIDN: 0525088601

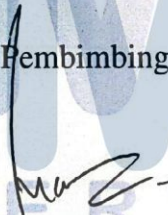
Penguji



(SY. Yuliani Yakub, S.Kom., M.T., PhD.,
CEH)

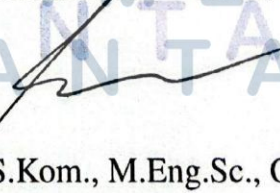
NIDN: 0411037904

Pembimbing



(Marlinda Vasty Overbeek, S.Kom., M.Kom)
NIDN: 0818038501

Ketua Program Studi Informatika,



(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)
NIDN: 0315109103

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Vaness Febraio
NIM : 00000055310
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Analisis Sentimen Komentar YouTube
Berbahasa Indonesia terkait Isu
Warga Negara Indonesia di Jepang
Menggunakan Model Hybrid
IndoBERT dan LSTM

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 22 Desember 2025

Yang menyatakan



Vaness Febraio

HALAMAN PERSEMBAHAN / MOTTO

"I strive with all my strength, and entrust the final result to Buddha"

Vaness Febraio



KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas berkat, rahmat, dan karunia-Nya, penulis dapat menyelesaikan skripsi yang berjudul “Analisis Sentimen Komentar YouTube Berbahasa Indonesia terkait Isu Warga Negara Indonesia di Jepang Menggunakan Model Hybrid IndoBERT dan LSTM” dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom.) pada Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara.

Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada :

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Ibu Marlinda Vasty Overbeek, S.Kom., M.Kom, sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.
6. Kepada Muhammad Rajja Farabi telah mendukung penulis untuk mengerjakan skripsi ini hingga selesai.

Penulis menyadari bahwa laporan skripsi ini masih jauh dari sempurna. Oleh karena itu, saran dan kritik yang membangun sangat penulis harapkan untuk perbaikan di masa mendatang. Semoga laporan ini dapat memberikan manfaat dan menjadi referensi bagi pembaca.

Tangerang, 22 Desember 2025



Vaness Febraio

ANALISIS SENTIMEN KOMENTAR YOUTUBE BERBAHASA INDONESIA TERKAIT ISU WARGA NEGARA INDONESIA DI JEPANG MENGUNAKAN MODEL HYBRID INDOBERT DAN LSTM

Vaness Febraio

ABSTRAK

Perkembangan pesat media sosial telah mengubah cara masyarakat mengekspresikan opini dan emosi terhadap berbagai isu sosial, dengan YouTube menjadi salah satu platform utama dalam diskursus publik. Salah satu isu yang menarik perhatian luas adalah keterlibatan Warga Negara Indonesia (WNI) dalam berbagai insiden di Jepang yang memicu beragam reaksi dari masyarakat Indonesia. Reaksi tersebut tercermin dalam komentar *YouTube* berbahasa Indonesia yang mengandung sentimen positif, negatif, dan netral. Penelitian ini bertujuan untuk menganalisis sentimen komentar *YouTube* terkait isu WNI di Jepang menggunakan model *hybrid IndoBERT-LSTM*. *IndoBERT* dimanfaatkan untuk menghasilkan *embedding kontekstual* yang mampu merepresentasikan makna semantik teks secara akurat, sedangkan *Long Short-Term Memory (LSTM)* digunakan untuk memodelkan urutan kata serta dependensi jangka panjang dalam kalimat. Dataset dikumpulkan dari platform YouTube dan melalui beberapa tahap praproses, meliputi *pembersihan teks*, *normalisasi*, *tokenisasi*, penghapusan *stopword*, dan *stemming*. Proses pelabelan sentimen dilakukan menggunakan *InSet Lexicon*, kemudian data dibagi menjadi data latih dan data uji dengan rasio 80:20. Hasil eksperimen menunjukkan bahwa model *IndoBERT-LSTM* mencapai akurasi sebesar 77%, dengan nilai *precision*, *recall*, dan *F1-score* masing-masing sebesar 0,77. Secara per kelas, sentimen negatif memperoleh *F1-score* 0,79, sentimen netral 0,71, dan sentimen positif 0,80. Selain itu, hasil analisis menunjukkan bahwa sentimen negatif mendominasi komentar, yang mencerminkan kuatnya respons emosional masyarakat terhadap isu tersebut serta menegaskan efektivitas model yang diusulkan dalam analisis sentimen teks media sosial berbahasa Indonesia.

Kata kunci: Analisis sentimen, *IndoBERT*, *LSTM*, Hibrida, *YouTube*

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

***SENTIMENT ANALYSIS OF INDONESIAN-LANGUAGE YOUTUBE
COMMENTS ON THE ISSUE OF INDONESIAN CITIZENS IN JAPAN
USING THE INDOBERT AND LSTM HYBRID MODEL***

Vaness Febraio

ABSTRACT

The rapid development of social media has changed the way people express opinions and emotions toward various social issues, with YouTube becoming one of the main platforms for public discourse. One issue that has attracted widespread attention is the involvement of Indonesian citizens (WNI) in various incidents in Japan, which has triggered diverse reactions from Indonesian society. These reactions are reflected in Indonesian-language YouTube comments containing positive, negative, and neutral sentiments. This study aims to analyze the sentiment of YouTube comments related to the issue of Indonesian citizens in Japan using a hybrid IndoBERT–LSTM model. IndoBERT is utilized to generate contextualized embeddings that accurately represent the semantic meaning of text, while Long Short-Term Memory (LSTM) is employed to model word sequences and long-term dependencies within sentences. The dataset was collected from YouTube and processed through several preprocessing stages, including text cleaning, normalization, tokenization, stopword removal, and stemming. Sentiment labeling was conducted using the InSet Lexicon, and the data were divided into training and testing sets with an 80:20 ratio. The experimental results show that the IndoBERT–LSTM model achieves an accuracy of 77%, with precision, recall, and F1-score values of 0.77. Class-wise analysis indicates that the negative sentiment class obtains an F1-score of 0.79, the neutral class 0.71, and the positive class 0.80. Furthermore, the results indicate that negative sentiment dominates the comments, reflecting strong emotional responses from the public and confirming the effectiveness of the proposed model for sentiment analysis of Indonesian-language social media text.

Keywords: *Sentiment analysis, IndoBERT, LSTM, Hybrid, Youtube*

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR.....	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xi
DAFTAR GAMBAR	xii
DAFTAR KODE.....	xiii
DAFTAR LAMPIRAN	xiv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	4
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
1.5.1 Manfaat Akademis	5
1.5.2 Manfaat Praktis	5
1.6 Sistematika Penulisan	5
BAB 2 LANDASAN TEORI.....	7
2.1 Sentiment Analysis	7
2.2 Lexicon Based Approach	7
2.2.1 InSet Lexicon	8
2.3 Natural Language Processing (NLP)	9
2.4 Deep Learning	9
2.5 Long Short Term Memories (LSTM).....	10
2.6 Transformers	16
2.6.1 Encoder	18
2.6.2 Input Embedding dan Output Embedding	21
2.6.3 Decoder	22
2.7 BERT	25
2.7.1 Tahapan Proses pada BERT	26
2.7.2 Aplikasi BERT pada Bahasa Indonesia	29
2.8 indobERT	30
2.9 Hyperparameter Tuning	32
2.10 Evaluasi	32
2.10.1 Metrik Evaluasi	33
BAB 3 METODOLOGI PENELITIAN	35
3.1 Metode Penelitian	35
3.2 Literature Review	36
3.3 Data Collection	36
3.4 Preprocessing Data	38
3.5 Data Labelling	39
3.6 Split Dataset.....	42
3.7 Hyperparameter Tuning.....	43

3.8	Modeling.....	44
3.8.1	Input Kalimat Teks.....	46
3.8.2	Tokenisasi Menggunakan IndoBERT.....	46
3.8.3	Ekstraksi Embedding Kontekstual	46
3.8.4	Pemodelan Urutan Menggunakan LSTM.....	46
3.8.5	Dense Layer dan Fungsi Softmax.....	47
3.8.6	Penentuan Label Sentimen	47
3.8.7	Contoh Proses Klasifikasi Kalimat.....	47
3.8.8	Parameter Pelatihan Model	47
3.9	Evaluation Model.....	48
BAB 4	HASIL DAN DISKUSI	50
4.1	Data Collection	50
4.1.1	Menentukan Video	50
4.1.2	Mengambil Data Video	50
4.1.3	Ekstraksi Komentar	51
4.1.4	Hasil disimpan ke CSV	54
4.2	Preprocessing Data	56
4.2.1	Case Folding.....	56
4.2.2	Cleaning Text	57
4.2.3	Normalization.....	57
4.2.4	Tokenization.....	59
4.2.5	Stopword Removal.....	60
4.2.6	Stemming	61
4.3	Data Labeling.....	62
4.3.1	InSet Lexicon	62
4.3.2	Proses Pelabelan Data	63
4.4	Split Dataset.....	66
4.5	Hyperparameter Tuning.....	67
4.5.1	Persiapan Library	68
4.5.2	Tokenisasi dan Encoding Data.....	68
4.5.3	Perancangan Model dan Ruang Hyperparameter	69
4.5.4	Membuat Model IndoBERT–LSTM	69
4.5.5	Proses Hyperparameter Tuning	70
4.6	Modeling.....	71
4.6.1	Arsitektur Model IndoBERT–LSTM	71
4.6.2	Implementasi Model.....	73
4.6.3	Hasil Pelatihan Model	75
4.7	Evaluasi Model	76
4.7.1	Implementasi Kode Evaluasi	77
4.7.2	Hasil Evaluasi Kinerja Model	78
4.8	Hasil Pengujian Model	80
BAB 5	SIMPULAN DAN SARAN.....	84
5.1	Saran.....	84
DAFTAR PUSTAKA		85

DAFTAR TABEL

Tabel 2.1	Contoh Kata dan score pada InSet Lexicon	9
Tabel 4.1	Hasil Pelatihan Model Hybrid IndoBERT-LSTM	75
Tabel 4.2	Classification Report Model IndoBERT-LSTM	80



DAFTAR GAMBAR

Gambar 2.1	Arsitektur LSTM	10
Gambar 2.2	Arsitektur Forget Gate pada LSTM	11
Gambar 2.3	Arsitektur Input Gate pada LSTM	12
Gambar 2.4	Arsitektur Output Gate pada LSTM	13
Gambar 2.5	Arsitektur Cell State pada LSTM	15
Gambar 2.6	Arsitektur model Transformers	17
Gambar 2.7	Arsitektur Encoder	18
Gambar 2.8	Mekanisme dari Attention	19
Gambar 2.9	Feed Forward Network	20
Gambar 2.10	Arsitektur Embedding	21
Gambar 2.11	Arsitektur Decoder	22
Gambar 2.12	Arsitektur Masked Multi-Head Attention pada Decoder	23
Gambar 2.13	Arsitektur Cross-Attention pada Decoder	24
Gambar 2.14	Arsitektur BERT	25
Gambar 2.15	Proses Embedding pada Model BERT	27
Gambar 2.16	Tahapan Pre-training dan Fine-tuning pada Model BERT	28
Gambar 3.1	Diagram alur penelitian	35
Gambar 3.2	Flowchart Proses Pengumpulan Data	37
Gambar 3.3	Preprocessing	38
Gambar 3.4	Labelling	40
Gambar 3.5	Hyperparameter Tuning	43
Gambar 3.6	Modeling	45
Gambar 3.7	Evaluation Model	48
Gambar 4.1	Hasil Ekstraksi	55
Gambar 4.2	Hasil distribusi komentar	55
Gambar 4.3	Hasil Case Folding	56
Gambar 4.4	Hasil Cleaning Text	57
Gambar 4.5	Hasil Normalization	58
Gambar 4.6	Hasil Persentase Normalization	59
Gambar 4.7	Hasil Tokenization	60
Gambar 4.8	Hasil Stopwords Removal	61
Gambar 4.9	Hasil Stemming	62
Gambar 4.10	Hasil Labeling	66
Gambar 4.11	Hasil Split Dataset	67
Gambar 4.12	Hasil Arsitektur model IndoBERT-LSTM	72
Gambar 4.13	Hasil Confusion Matrix	79
Gambar 4.14	Hasil uji model	82

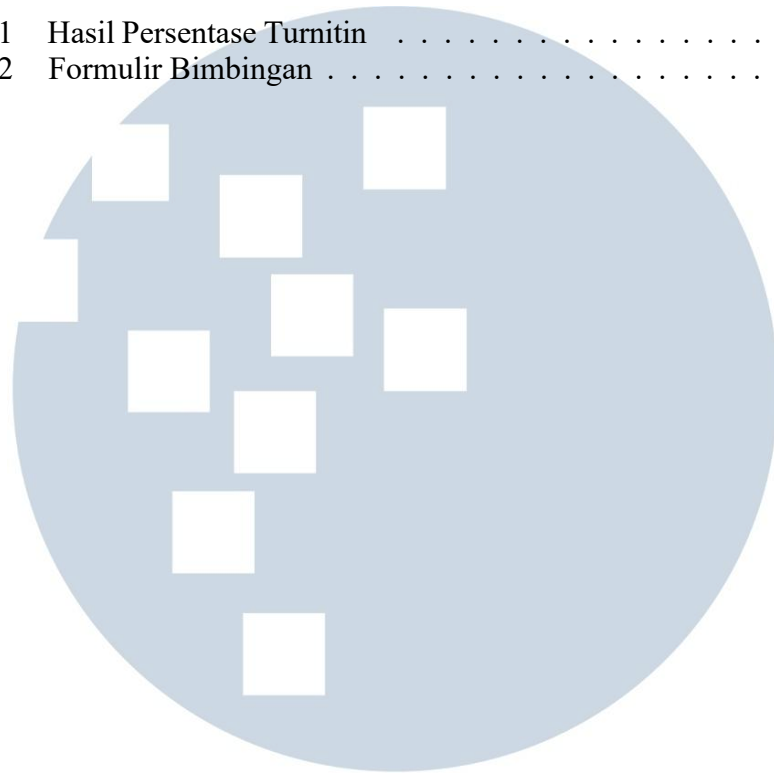
DAFTAR KODE

Kode 2.1	Pseudocode Indobert	30
Kode 3.1	Pseudocode Split Dataset	42
Kode 4.1	Kumpulan data API	50
Kode 4.2	ekstraksi komentar parent	51
Kode 4.3	ekstraksi komentar balasan	52
Kode 4.4	Proses ekstraksi komentar	54
Kode 4.5	proses penyimpanan data komentar ke dalam csv	55
Kode 4.6	ekstraksi komentar parent	55
Kode 4.7	proses case folding	56
Kode 4.8	proses cleaning text	57
Kode 4.9	proses normalization	57
Kode 4.10	cek jumlah slang	58
Kode 4.11	persentase keberhasilan normalisasi	59
Kode 4.12	proses tokenization	59
Kode 4.13	proses Stopwords removal	60
Kode 4.14	proses stemming	61
Kode 4.15	proses labeling	63
Kode 4.16	kode untuk menampilkan hasil labeling	64
Kode 4.17	proses split data	66
Kode 4.18	Library yang akan digunakan	68
Kode 4.19	Tokenisasi dan encoding	68
Kode 4.20	Pencarian model	69
Kode 4.21	Model IndoBERT-LSTM	69
Kode 4.22	Hyperparameter Tuning	70
Kode 4.23	Kode untuk Pemilihan Konfigurasi terbaik	73
Kode 4.24	Kode untuk ekstraksi hyperparameter terbaik	74
Kode 4.25	Kode untuk pembangunan model	74
Kode 4.26	Kode untuk pelatihan model	74
Kode 4.27	proses evaluasi model	77
Kode 4.28	Kalimat uji coba	81
Kode 4.29	Uji coba model	81

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin	89
Lampiran 2	Formulir Bimbingan	91



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA