

**ANALISIS SENTIMEN MASYARAKAT INDONESIA
TERHADAP BERITA MEGA KORUPSI
PERTAMINA MENGGUNAKAN
ALGORITMA INDOBERT**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

ILLYAZ ARYA SHIHAB KUSUMA

22.11.5156

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

**ANALISIS SENTIMEN MASYARAKAT INDONESIA
TERHADAP BERITA MEGA KORUPSI
PERTAMINA MENGGUNAKAN
ALGORITMA INDOBERT**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

ILLYAZ ARYA SHIHAB KUSUMA

22.11.5156

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS SENTIMEN MASYARAKAT INDONESIA
TERHADAP BERITA MEGA KORUPSI
PERTAMINA MENGGUNAKAN
ALGORITMA INDOBERT**

yang disusun dan diajukan oleh

Illyaz Arya Shihab Kusuma
22.11.5156

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 28 Januari 2026

Dosen Pembimbing,



Ike Verawati, M.Kom
NIK. 190302237

HALAMAN PENGESAHAN

SKRIPSI

ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP BERITA MEGA KORUPSI PERTAMINA MENGGUNAKAN ALGORITMA INDOBERT

yang disusun dan diajukan oleh

Illyaz Arya Shihab Kusuma
22.11.5156

Telah dipertahankan di depan Dewan Penguji
pada tanggal 28 Januari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Ali Mustopa, S.Kom, M.Kom
NIK. 190302192

Ahmad Sa'di, S.Kom, M.Eng
NIK. 190302459

Ike Verawati, S.Kom, M.Kom
NIK. 190302237



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 28 Januari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Illyaz Arya Shihab Kusuma
NIM : 22.11.5156

Menyatakan bahwa Skripsi dengan judul berikut:

ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP BERITA MEGA KORUPSI PERTAMINA MENGGUNAKAN ALGORITMA INDOBERT

Dosen Pembimbing : Ike Verawati, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 28 Januari 2026

Yang Menyatakan,



Illyaz Arya Shihab Kusuma

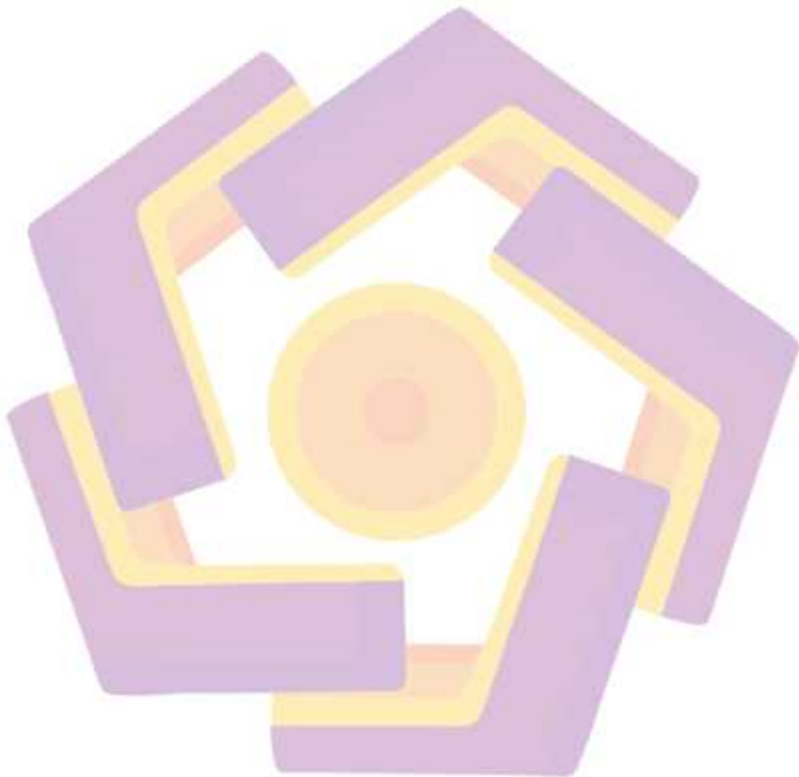
HALAMAN PERSEMBAHAN

Dengan segenap rasa syukur kepada Tuhan Yang Maha Esa atas limpahan rahmat, karunia, dan kekuatan yang senantiasa menyertai dalam setiap langkah, penulis dapat menyelesaikan skripsi ini sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di Universitas Amikom Yogyakarta. Segala proses yang dilalui selama penyusunan skripsi ini merupakan perjalanan yang penuh dengan pembelajaran, tantangan, dan pengalaman berharga.

Sebagai wujud rasa hormat dan terima kasih yang mendalam, penulis mempersembahkan karya ini kepada:

1. Orang tua tercinta, yang telah menjadi sumber kekuatan terbesar dalam hidup penulis. Terima kasih atas kasih sayang, doa, dukungan moral dan materil yang tak terhingga. Tanpa keikhlasan dan pengorbanan kalian, penulis tidak akan mampu sampai pada tahap ini.
2. Adik-adik penulis, yang selalu menjadi penyemangat dan sumber inspirasi, serta turut memberikan dukungan dan keceriaan di setiap waktu. Kehadiran kalian menjadi pengingat akan tanggung jawab dan harapan untuk menjadi panutan yang baik.
3. Dosen pembimbing, yang telah memberikan waktu, perhatian, arahan, serta masukan yang membangun selama proses penyusunan skripsi ini. Terima kasih atas kesabaran dan dedikasinya dalam membimbing hingga skripsi ini terselesaikan dengan baik.
4. Seluruh dosen Universitas Amikom Yogyakarta, yang telah memberikan ilmu, wawasan, dan pengalaman akademik selama masa studi. Semoga segala ilmu yang diberikan menjadi bekal yang bermanfaat untuk masa depan.
5. Teman-teman seperjuangan, yang telah menjadi bagian dari perjalanan ini, baik dalam suka maupun duka. Terima kasih atas semangat, kerja sama, dan kebersamaan yang tidak ternilai harganya. Semoga perjuangan kita semua membuahkan hasil yang membanggakan.

6. Semua pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan bantuan, dukungan, serta doa dalam bentuk apa pun. Setiap bantuan sangat berarti dan akan selalu dikenang dengan penuh rasa terima kasih.



KATA PENGANTAR

Segala puji dan syukur penulis panjatkan atas kehadiran Tuhan Yang Maha Esa atas rahmat dan berkah-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul: "ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP BERITA MEGA KORUPSI PERTAMINA MENGGUNAKAN ALGORITMA INDOBERT".

Penyusunan skripsi ini tentunya tidak terlepas dari dukungan, bimbingan, dan bantuan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada seluruh pihak yang terlibat.

Penulis menyadari bahwa skripsi ini jauh dari kata sempurna. Dalam pembuatannya, masih banyak kesalahan maupun kekurangan. Oleh karena itu, penulis sangat terbuka terhadap segala kritik dan saran yang membangun demi menyempurnakan skripsi ini. Namun, penulis tetap berharap semoga skripsi ini tetap bermanfaat bagi pembaca.

Yogyakarta, 25 Januari 2026

Penulis



Illyaz Arya Shihab Kusuma

DAFTAR ISI

HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR.....	vii
DAFTAR TABEL	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMPIRAN.....	xii
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH	xv
INTISARI	xi
ABSTRACT	xx
BAB I.....	1
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian.....	3
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Studi Literatur.....	5
2.2 Dasar Teori.....	18
2.2.1 Media Sosial dan Twitter	18
2.2.2 Analisis Sentimen	18
2.2.3 Natural Language Processing (NLP)	18
2.2.4 Transformer dan BERT.....	19
2.2.5 IndoBERT	19
2.2.6 Metrik Evaluasi Model Klasifikasi	22
BAB III.....	24
METODE PENELITIAN	24
3.1 Objek Penelitian	24
3.2 Alur Penelitian.....	24

3.2.1	Scraping Twitter dan TikTok	25
3.2.2	Merge Dataset	25
3.2.3	Preprocessing	26
3.2.4	Fine-tuning IndoBERT	28
3.2.5	Sentiment Classification	31
3.2.6	Evaluasi Model	31
3.3	Alat dan Bahan	31
BAB IV		34
HASIL DAN PEMBAHASAN		34
4.1	Scraping Data	34
4.2	Pre-processing Data	36
4.2.1	Penggabungan Data (Data Merging)	37
4.2.2	Pembersihan Data (Data Cleaning)	37
4.2.3	Case Folding	42
4.2.4	Tokenisasi (Tokenization)	44
4.2.5	Normalisasi Kata (Word Normalization)	45
4.2.6	Validasi dan Pembersihan Duplikasi Data	47
4.3	Pelabelan Data	48
4.4	Exploratory Data Analysis (EDA)	50
4.4.1	Distribusi Kelas Sentimen	50
4.4.2	Visualisasi Word Cloud	51
4.4.3	Distribusi Panjang Kalimat (Jumlah Token)	52
4.5	Pembagian Data (Data Splitting)	53
4.6	Implementasi Fine Tuning Model IndoBERT Dengan Custom Trainer	54
4.7	Evaluasi dan Analisis Hasil	57
4.7.1	Confusion Matrix	58
4.7.2	Laporan Klasifikasi	58
BAB V		60
KESIMPULAN DAN SARAN		60
5.1	Kesimpulan	60
5.2	Saran	61
DAFTAR PUSTAKA		62

DAFTAR TABEL

Tabel 1. 1 Keaslian Penulisan	8
Tabel 3. 1 Dataset.....	32
Tabel 4. 1 Hasil Scraping Twitter	35
Tabel 4. 2 Hasil Scraping TikTok	35
Tabel 4. 3 Data Sebelum Tahap Pembersihan	36
Tabel 4. 4 Penggabungan Data	37
Tabel 4. 5 Tahap Menghapus User Mention.....	38
Tabel 4. 6 Tahap Menghapus URL	39
Tabel 4. 7 Menghapus Angka dan Simbol.....	40
Tabel 4. 8 Penghapus Karakter Tunggal	41
Tabel 4. 9 Case Folding	43
Tabel 4. 10 Tokenisasi Teks	44
Tabel 4. 11 Normalisasi Kata.....	46
Tabel 4. 12 Pre-processing Data	47
Tabel 4. 13 Pelabelan Data	49
Tabel 4. 14 Konfigurasi Hiperparameter	54
Tabel 4. 15 Perbandingan Hasil Lexicon vs. IndoBERT	55
Tabel 4. 16 Hasil Pengukuran Kuantitatif.....	59

DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian	25
Gambar 4. 1 Hasil Distribusi Sentimen Lexicon	51
Gambar 4. 2 Visualisasi Word Cloud	52
Gambar 4. 3 Distribusi Panjang Kata.....	53
Gambar 4. 4 Perbandingan Distribusi Sentimen: Lexicon vs IndoBERT.....	57
Gambar 4. 5 Confusion Matrix.....	58



DAFTAR LAMPIRAN

Lampiran 1. Data Scrapping X (Twitter).....	65
Lampiran 2. Dataset Scrapping TikTok.....	67
Lampiran 3. Perbandingan Sebelum dan Sesudah <i>Preprocessing</i>	68
Lampiran 4. Perbandingan Hasil Klasifikasi Lexicon Dengan IndoBERT	73
Lampiran 5. Grafik Training History.....	77
Lampiran 6. Perbandingan Hasil Klasifikasi Lecixon Dengan IndoBERT	77



DAFTAR LAMBANG DAN SINGKATAN



BUMN	Badan Usaha Milik Negara
FN	False Negative
FP	False Positive
GPU	Graphic Processing Unit
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IndoBERT	Indonesian BERT
KBBI	Kamus Besar Bahasa Indonesia
KPK	Komisi Pemberantasan Korupsi
LR	Learning Rate
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OOV	Out-Of-Vocabulary
RAM	Random Access Memory
Regex	Regular Expression
SaaS	Software as a Service
TN	True Negative
TP	True Positive
URL	Uniform Resource Locator
Acc	Accuracy (Akurasi model)
P	Precision (Presisi)
R	Recall (Sensitivitas)
Σ	Sigma (Notasi penjumlahan, sering muncul di rumus akurasi/loss)
x	Input Vektor (Representasi teks input)
y	Label Target (Sentiment asli)
$\hat{y}$	Label Prediksi (Sentimen hasil prediksi)

V (Output akhir)

Penjumlahan berbobot dari *Value* berdasarkan bobot

Q

Query (Vektor pencari informasi)

K

Key (Vektor kunci informasi untuk dicocokkan)

V

Value (Vektor isi informasi untuk diambil)

W

Matriks bobot (*weights*) untuk lapisan linear pertama

b

Vektor bias untuk lapisan linear pertama

σ

Varians elemen x (di dimensi fitur yang sama)

β

Konstanta kecil untuk stabilitas numèrik (hindari bagi 0)

ϵ

Rata-rata elemen x (dihitung di dimensi fitur, per token)

μ

Persen

%

Kurang Dari

<

Lebih Dari

>

Sama Dengan

=

Bobot kelas (*Class Weight*)

w_j

Skor harmonik rata-rata presisi dan *recall* (*F1-Score*)

$F1$

Jumlah total data

N

DAFTAR ISTILAH

Analisis Sentimen

Proses komputasi untuk mengidentifikasi dan mengategorikan opini dalam teks, guna menentukan apakah sikap penulis terhadap topik tertentu adalah positif, negatif, atau netral.

Attention Mechanism

Mekanisme dalam arsitektur Transformer yang memungkinkan model untuk fokus pada bagian kalimat yang paling relevan saat memproses kata tertentu.

Batch Size

Jumlah sampel data pelatihan yang diproses oleh model dalam satu iterasi sebelum memperbarui parameter internal model.

Class Weighting

Teknik penanganan data tidak seimbang dengan memberikan bobot penalti lebih besar pada kelas minoritas agar model lebih memperhatikannya.

Confusion Matrix

Tabel evaluasi model klasifikasi yang membandingkan prediksi model dengan kebenaran dasar, terdiri dari *True Positive*, *True Negative*, *False Positive*, dan *False Negative*.

Deep Learning

Cabang pembelajaran mesin (*Machine Learning*) yang menggunakan jaringan saraf tiruan berlapis (*layers*) untuk mempelajari representasi data yang kompleks.



Epoch	Satu siklus lengkap di mana seluruh dataset pelatihan telah melewati proses pembelajaran (maju dan mundur) oleh model.
Fine-Tuning	Proses melatih kembali model pra-latih (<i>pre-trained</i>) dengan dataset baru yang lebih spesifik untuk mengoptimalkan performa pada tugas tertentu.
Hyperparameter	Parameter konfigurasi eksternal yang nilainya ditetapkan secara manual sebelum pelatihan dimulai (contoh: <i>learning rate</i> , <i>epoch</i>) dan tidak dipelajari dari data.
Imbalanced Data	Kondisi dalam dataset klasifikasi di mana jumlah sampel pada satu kelas jauh lebih dominan dibandingkan kelas lainnya.
IndoBERT	Model bahasa berbasis arsitektur BERT yang telah dilatih secara khusus menggunakan korpus data teks Bahasa Indonesia dalam skala besar.
Learning Rate	Parameter yang mengontrol kecepatan model dalam belajar dengan menentukan seberapa besar perubahan bobot model pada setiap langkah pembaruan.
Loss Function	Fungsi matematis yang mengukur selisih antara prediksi model dengan target sebenarnya; tujuannya adalah meminimalkan nilai ini selama pelatihan.

Natural Language Processing

Cabang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa alami manusia, mencakup pemahaman dan manipulasi teks.

Overfitting

Kondisi di mana model terlalu menghafal data latih hingga detail *noise*-nya, sehingga performanya buruk saat diuji pada data baru.

Pre-trained Model

Model yang telah dilatih sebelumnya pada dataset besar untuk menyelesaikan masalah umum, dan dapat digunakan kembali sebagai dasar tugas lain.

Scraping

Teknik ekstraksi data secara otomatis dari situs web atau platform media sosial menggunakan program komputer.

Tokenisasi

Proses memecah teks atau kalimat menjadi unit-unit kecil (token) agar dapat diproses oleh komputer; pada IndoBERT menggunakan metode *WordPiece*.

Transformer

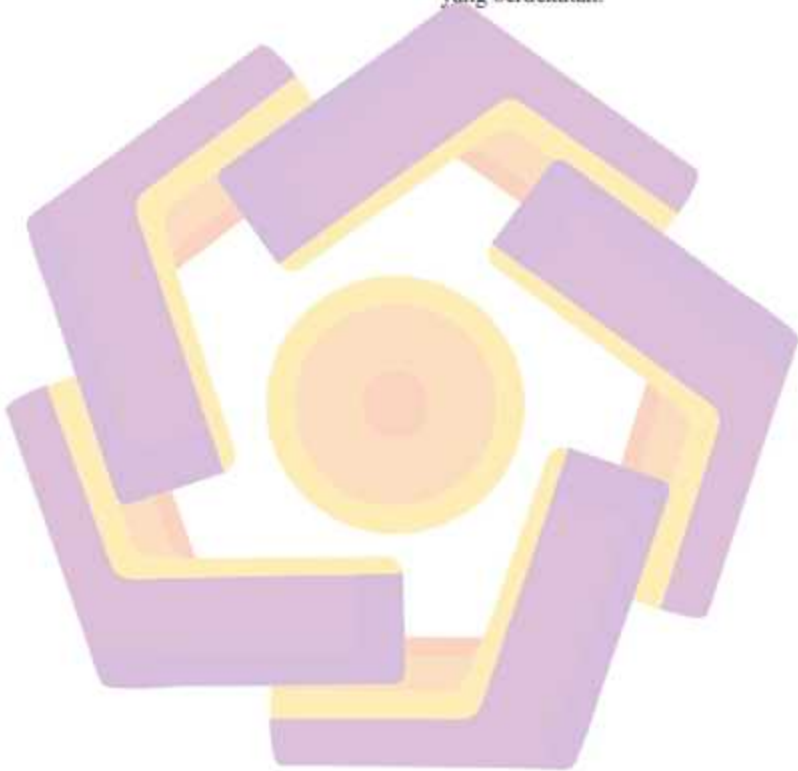
Arsitektur *Deep Learning* yang mengandalkan mekanisme *self-attention* untuk memproses data teks secara paralel, menjadi dasar dari model BERT.

Vektor

Besaran matematis yang digunakan untuk merepresentasikan fitur teks

Word Embedding

(kata/kalimat) dalam ruang dimensi tinggi agar dapat dihitung oleh model. Representasi kata dalam bentuk vektor bilangan riil di mana kata yang bermakna serupa memiliki jarak vektor yang berdekatan.



INTISARI

Berita mengenai Pertamina sebagai salah satu Badan Usaha Milik Negara (BUMN) sering kali menjadi perhatian publik, khususnya di media sosial. Opini masyarakat yang terekam dalam bentuk komentar dan unggahan dapat merepresentasikan sentimen masyarakat, baik positif maupun negatif. Ketika opini tersebut tidak dianalisis secara tepat, maka dapat menimbulkan mispersepsi yang berdampak pada citra perusahaan.

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap berita mengenai Pertamina menggunakan pendekatan pemrosesan bahasa alami (Natural Language Processing) dengan algoritma IndoBERT. Data dikumpulkan dari berbagai sumber berita daring dan media sosial, kemudian dilakukan tahap praproses seperti penghapusan duplikat, konversi huruf kecil, normalisasi kata tidak baku, pemisahan huruf alfabet dari non-alfabet, serta validasi teks untuk menghindari kesalahan dalam tokenisasi.

Proses analisis dilakukan menggunakan pipeline model IndoBERT untuk klasifikasi sentimen secara otomatis. Hasil penelitian menunjukkan bahwa IndoBERT mampu mengklasifikasikan sentimen masyarakat terhadap berita Pertamina dengan akurasi yang baik setelah melalui tahapan praproses yang optimal. Penelitian ini memberikan kontribusi dalam mendukung pengambilan keputusan strategis perusahaan berdasarkan persepsi publik.

Hasil penelitian ini bermanfaat bagi institusi pemerintah, pelaku industri, dan peneliti di bidang data sains serta dapat dijadikan acuan dalam mengembangkan sistem pemantauan opini publik secara real-time. Penelitian lanjutan disarankan untuk mengeksplorasi model lain atau data dengan cakupan waktu yang lebih luas.

Kata kunci: IndoBERT, NLP, Opini publik, Pertamina, Sentimen

ABSTRACT

News about Pertamina, as one of Indonesia's state-owned enterprises (BUMN), often garners public attention, especially on social media. Public opinions expressed through comments and posts can reflect societal sentiment, whether positive or negative. When such opinions are not properly analyzed, they may lead to misperceptions that could harm the company's reputation.

This study aims to analyze the sentiments of Indonesian society toward news about Pertamina using Natural Language Processing (NLP) techniques with the IndoBERT algorithm. The data was collected from various online news sources and social media platforms, followed by preprocessing steps such as duplicate removal, lowercasing, normalization of non-standard words, separation of alphabets from non-alphabet characters, and text validation to prevent tokenization errors.

The sentiment analysis process was carried out using the IndoBERT pipeline model for automatic sentiment classification. The results indicate that IndoBERT can effectively classify public sentiment regarding Pertamina news with good accuracy after optimal preprocessing. This research contributes to supporting strategic corporate decision-making based on public perception.

The results are beneficial for government institutions, industry stakeholders, and researchers in the field of data science and can serve as a reference for developing real-time public opinion monitoring systems. Further studies are recommended to explore alternative models or datasets with broader time coverage.

Keywords: IndoBERT, NLP, Pertamina, public opinion, Sentiment