

**IMPLEMENTASI *NATURAL LANGUAGE PROCESSING* UNTUK  
*LABELING* DATASET ULASAN *ATTRACTION* DI INDONESIA**

**TUGAS AKHIR**



**FENESIA SAMOSIR  
1212001029**

**PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNIK DAN ILMU KOMPUTER  
UNIVERSITAS BAKRIE  
JAKARTA**

**2026**

**IMPLEMENTASI *NATURAL LANGUAGE PROCESSING* UNTUK  
*LABELING* DATASET ULASAN *ATTRACTION* DI INDONESIA**

**TUGAS AKHIR**

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



**FENESIA SAMOSIR  
1212001029**

**PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNIK DAN ILMU KOMPUTER  
UNIVERSITAS BAKRIE  
JAKARTA**

**2026**

## **HALAMAN PERNYATAAN ORISINALITAS**

Tugas Akhir ini adalah hasil karya saya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : Fenesia Samosir

NIM : 1212001029

Tanda Tangan :



Tanggal : 16 September 2026

## HALAMAN PENGESAHAN

Tugas Akhir ini diajukan oleh:

Nama : Fenesia Samosir  
NIM : 1212001029  
Program Studi : Informatika  
Fakultas : Teknik dan Ilmu Komputer  
Judul Skripsi : “IMPLEMENTASI *NATURAL LANGUAGE PROCESSING*  
UNTUK *LABELING* DATASET ULASAN *ATTRACTION*  
DI INDONESIA”

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer (S.Kom) pada Program Studi Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Bakrie.

## DEWAN PENGUJI

Pembimbing 1 : Guson P. Kuntarto, S.T., M.Sc.

(  )  
GP 18/08/2026 11:38 WIB

Pembimbing 2 : Irwan Prasetya Gunawan, S.T., M.Eng., Ph.D.

(  )  
2026-09-18 18:06:47

Penguji 1 : Dimas Aryo Anggoro, S.Kom., M.Sc.

(  )  
18/09/2026 9:02 WIB

Signed with  privy

Penguji 2 : Berkah Iman Santoso, ST, MTI

(  )

B. I. S. (LOBER4857) - September 17, 2026 - 18:32:02

Ditetapkan di : Jakarta

Tanggal : 16 September 2026

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai sivitas akademik Universitas Bakrie, saya yang bertanda tangan di bawah ini:

Nama : Fenesia Samosir  
NIM : 1212001029  
Program Studi : Informatika  
Fakultas : Teknik dan Ilmu Komputer  
Jenis Tugas Akhir : Tugas Akhir

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Bakrie **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul:

**Implementasi *Natural Language Processing* (NLP) untuk *Labeling* Dataset Ulasan *Attraction* di Indonesia**

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Bakrie berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta untuk kepentingan akademis.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Pada tanggal : 16 September 2026

Yang Menyatakan



( Fenesia Samosir )

## KATA PENGANTAR

Puji syukur saya panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga saya dapat menyusun Tugas Akhir ini yang berjudul "*Implementasi Natural Language Processing untuk Labeling Dataset Ulasan Attraction di Indonesia.*" Tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Bakrie. Dalam proses penyusunan tugas akhir ini, saya banyak mendapatkan bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Ibunda tercinta, yang meskipun telah tiada sejak saya masih berusia 15 tahun, namun kasih sayangnya tetap hidup dan abadi di hati saya, menjadi sumber kekuatan yang tak tergantikan dalam setiap langkah hidup saya. Serta Ayah, Kak Martha, Abang Shem, dan seluruh keluarga saya, atas doa, dukungan moral, dan semangat yang tiada henti, yang senantiasa menguatkan saya dalam menyelesaikan studi ini.
2. Ibu Prof. Ir. Sofia W. Alisjahbana, M.Sc., Ph.D., IPU., ASEAN Eng. Selaku rektor Universitas Bakrie.
3. Bapak Dr. Mohammad Ihsan, S.T., M.Sc. selaku Dekan Fakultas Teknik dan Ilmu Komputer.
4. Bapak Iwan Adhicandra, S.T., M.Sc, MIEEE, MACM. Selaku Ketua Program Studi Teknik Informatika Universitas Bakrie.
5. Bapak Guson P. Kuntarto, S.T., M. Sc. Selaku dosen pembimbing pertama saya yang telah memberikan bimbingan, arahan, dan dukungan selama proses penelitian dan penyusunan tugas akhir ini.
6. Bapak Irwan Prasetya Gunawan, S.T., M.Eng., Ph.D. selaku dosen pembimbing kedua saya yang telah membantu saya dalam proses penyusunan tugas akhir ini.

7. Seluruh dosen Universitas Bakrie, khususnya di Program Studi Informatika, atas ilmu, bimbingan, serta dedikasi yang telah diberikan selama masa perkuliahan.
8. Teman-teman Hashira: Ardy, Fazrah, Gial, Malik, Hakim, dan Randy. Terima kasih atas tawa, semangat, dan kebersamaan selama masa perkuliahan.
9. Nofita, teman yang selalu hadir dalam suka dan duka, yang dengan sabar membantu saya saat sakit, memberi semangat, dan menjadi tempat berbagi selama proses penyusunan tugas akhir ini.
10. Kak Gaby yang sudah bersedia menemani saya dan senantiasa memberikan dukungan kepada saya dalam proses penyusunan tugas akhir ini.
11. Seluruh pihak yang telah memberikan dukungan, bantuan, dan doa selama proses penyusunan tugas akhir ini, yang tidak dapat disebutkan satu per satu.

Besar harapan saya, tugas akhir ini dapat bermanfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang *Natural Language Processing* di Indonesia.

Jakarta, 16 September 2026



Fenesia Samosir

1212001029

## ABSTRAK

### IMPLEMENTASI *NATURAL LANGUAGE PROCESSING* UNTUK *LABELING* DATASET ULASAN *ATTRACTION* DI INDONESIA

Fenesia Samosir

1212001029@student.bakrie.ac.id

---

Ketersediaan dataset ulasan wisata berlabel sentimen masih sangat terbatas, khususnya untuk ulasan *attraction* di Bali yang bersifat multibahasa. Kondisi ini menjadi hambatan dalam pengembangan sistem analisis sentimen yang mendukung pengelolaan pariwisata berbasis data. Penelitian ini bertujuan menghasilkan dataset ulasan *attraction* di Bali yang berlabel sentimen, serta mengevaluasi dan membandingkan lima pendekatan model *Natural Language Processing* untuk menentukan model paling efektif dalam klasifikasi sentimen multibahasa. Secara metodologis, penelitian ini dilaksanakan melalui tahapan mulai dari penyusunan dataset hingga evaluasi model klasifikasi sentimen. Data dikumpulkan melalui *web scraping* dari TripAdvisor dan Google Review, mencakup ulasan berbahasa Inggris, Indonesia, dan Spanyol. Dari 22.128 ulasan awal, proses pembersihan, verifikasi bahasa otomatis, dan praproses menghasilkan 10.777 ulasan. Label awal dibentuk melalui pendekatan proxy berbasis rating, dengan rating 1–2 dipetakan sebagai negatif, rating 3 sebagai netral, dan rating 4–5 sebagai positif. Dataset dibagi dengan rasio 70:15:15 menggunakan *stratified sampling* berdasarkan kombinasi bahasa dan sentimen. Lima model dievaluasi pada *test set* yang sama, yaitu XLM-RoBERTa *zero-shot*, mBERT-XNLI *zero-shot*, XLM-RoBERTa *fine-tuned*, mBERT *fine-tuned*, dan GPT *prompt-based zero-shot*, dengan *macro F1-score* sebagai metrik utama karena distribusi kelas tidak seimbang. GPT *prompt-based zero-shot* memperoleh *macro F1-score* tertinggi sebesar 70,88%, diikuti XLM-RoBERTa *fine-tuned* sebesar 70,74%. Namun XLM-RoBERTa *fine-tuned* unggul pada *accuracy* sebesar 89,98% dan *weighted F1-score* sebesar 89,68%, serta lebih layak diterapkan pada keseluruhan data secara mandiri. Model tersebut ditetapkan sebagai model terbaik dan digunakan untuk melabeli seluruh dataset, menghasilkan 10.777 ulasan berlabel sentimen dengan tingkat kesesuaian 93,28% terhadap proxy label sebagai luaran akhir penelitian.

**Kata kunci:** NLP, BERT, XLM-RoBERTa, klasifikasi sentimen, proxy label, dataset wisata, TripAdvisor, Google Review, Bali

## **ABSTRACT**

### **IMPLEMENTATION OF NATURAL LANGUAGE PROCESSING FOR LABELING ATTRACTION REVIEW DATASETS IN INDONESIA**

Fenesia Samosir

1212001029@student.bakrie.ac.id

---

*The availability of sentiment-labeled tourism review datasets remains limited, particularly for multilingual attraction reviews in Bali. This condition hinders the development of sentiment analysis systems supporting data driven tourism management. This study aims to produce a sentiment-labeled dataset of attraction reviews in Bali, and to evaluate and compare five Natural Language Processing approaches in determining the most effective model for multilingual sentiment classification. Methodologically, this study was carried out through stages ranging from dataset construction to the evaluation of the sentiment classification models. Data were collected through web scraping from TripAdvisor and Google Review, covering reviews in English, Indonesian, and Spanish. From 22,128 initial reviews, cleaning, automatic language verification, and preprocessing resulted in 10,777 reviews. Initial labels were formed using a rating-based proxy approach, mapping ratings of 1–2 as negative, 3 as neutral, and 4–5 as positive. The dataset was split at a ratio of 70:15:15 using stratified sampling based on the combination of language and sentiment. Five models were evaluated on the same test set, namely XLM-RoBERTa zero-shot, mBERT-XNLI zero-shot, XLM-RoBERTa fine-tuned, mBERT fine-tuned, and GPT prompt-based zero-shot, with macro F1-score as the primary metric due to the imbalanced class distribution. GPT prompt-based zero-shot achieved the highest macro F1-score of 70.88%, followed by XLM-RoBERTa fine-tuned at 70.74%. However, XLM-RoBERTa fine-tuned excelled in accuracy at 89.98% and weighted F1-score at 89.68%, and was more feasible to apply independently across the entire dataset. This model was therefore designated as the best model and used to label the whole dataset, producing 10,777 sentiment-labeled reviews with an agreement rate of 93.28% against the proxy labels as the final research output.*

**Keywords:** *NLP, BERT, XLM-RoBERTa, sentiment classification, proxy label, tourism dataset, TripAdvisor, Google Review, Bali*

## DAFTAR ISI

|                                                                                  |          |
|----------------------------------------------------------------------------------|----------|
| HALAMAN PERNYATAAN ORISINALITAS                                                  | ii       |
| HALAMAN PENGESAHAN                                                               | iii      |
| HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI                                         | iv       |
| KATA PENGANTAR                                                                   | v        |
| ABSTRAK                                                                          | vii      |
| DAFTAR ISI                                                                       | ix       |
| DAFTAR TABEL                                                                     | xiii     |
| DAFTAR GAMBAR                                                                    | xv       |
| DAFTAR SINGKATAN                                                                 | xvi      |
| DAFTAR ISTILAH                                                                   | xvii     |
| DAFTAR LAMPIRAN                                                                  | xxii     |
| <b>BAB I    PENDAHULUAN</b>                                                      | <b>1</b> |
| 1.1    Latar Belakang                                                            | 1        |
| 1.2    Rumusan Masalah                                                           | 3        |
| 1.3    Batasan Masalah                                                           | 4        |
| 1.4    Tujuan Penelitian                                                         | 4        |
| 1.5    Manfaat Penelitian                                                        | 5        |
| 1.6    Sistematika Penelitian                                                    | 5        |
| 1.7    Penutup                                                                   | 6        |
| <b>BAB II   TINJAUAN PUSTAKA</b>                                                 | <b>7</b> |
| 2.1    Tinjauan Literatur                                                        | 7        |
| 2.2 <i>E-Tourism</i>                                                             | 11       |
| 2.3    Teknik Dataset                                                            | 13       |
| 2.4 <i>Natural Language Processing</i>                                           | 14       |
| 2.4.1    Komponen dalam NLP                                                      | 14       |
| 2.4.2    Peran NLP dalam Proses <i>Labeling dan</i> Klasifikasi Sentimen         | 16       |
| 2.4.3    Klasifikasi NLP                                                         | 16       |
| 2.5    Model NLP <i>Modern</i>                                                   | 18       |
| 2.5.1 <i>Transformer</i>                                                         | 19       |
| 2.5.2    BERT ( <i>Bidirectional Encoder Representations from Transformers</i> ) | 22       |
| 2.5.3    mBERT                                                                   | 25       |
| 2.5.4    XLM-RoBERTa                                                             | 26       |
| 2.5.5 <i>Zero-Shot-Learning</i>                                                  | 26       |

|         |                                                               |    |
|---------|---------------------------------------------------------------|----|
| 2.5.6   | <i>Fine-Tuning</i>                                            | 27 |
| 2.5.7   | <i>GPT Prompt-Based Zero-Shot</i>                             | 27 |
| 2.6     | Proxy Label Berbasis Rating                                   | 28 |
| 2.7     | <i>Framework</i>                                              | 29 |
| 2.8     | Metrik Evaluasi                                               | 30 |
| 2.8.1   | Metrik Evaluasi Model                                         | 30 |
| 2.8.2   | Kriteria Sukses                                               | 32 |
| 2.9     | Distribusi Data Tidak Seimbang dan Strategi Pembagian Dataset | 33 |
| 2.9.1   | <i>Imbalanced Data</i>                                        | 33 |
| 2.9.2   | <i>Stratified Sampling</i>                                    | 34 |
| 2.10    | Penutup                                                       | 34 |
| BAB III | METODOLOGI PENELITIAN                                         | 36 |
| 3.1     | Kontribusi Penelitian                                         | 36 |
| 3.2     | Kerangka Penelitian                                           | 37 |
| 3.2.1   | Studi Literatur                                               | 38 |
| 3.2.2   | <i>Research Question</i> (Identifikasi Masalah)               | 39 |
| 3.2.3   | Pengumpulan dan Penyusunan Dataset                            | 39 |
| 3.2.4   | <i>Pre-Processing</i> Teks                                    | 46 |
| 3.2.5   | Pembentukan Proxy Label berbasis Rating                       | 47 |
| 3.2.6   | Pembagian Dataset                                             | 48 |
| 3.2.7   | Implementasi Model Klasifikasi Sentimen                       | 49 |
| 3.2.8   | Evaluasi dan Perbandingan Model                               | 52 |
| 3.2.9   | Penulisan Laporan                                             | 53 |
| 3.3     | Struktur Data / Data Model                                    | 53 |
| 3.4     | Instrumen Penelitian                                          | 55 |
| 3.5     | Penutup                                                       | 56 |
| BAB IV  | HASIL DAN PEMBAHASAN                                          | 58 |
| 4.1     | Hasil Penyusunan Dataset Penelitian                           | 58 |
| 4.1.1   | Hasil Akuisisi Data Referensi dan Ulasan Wisata               | 58 |
| 4.1.2   | Deskripsi Dataset Awal Hasil <i>Scraping</i>                  | 62 |
| 4.1.3   | Hasil Proses Cleaning Data Awal                               | 63 |
| 4.1.4   | Hasil Deteksi dan Normalisasi Bahasa                          | 63 |
| 4.1.5   | Hasil <i>Preprocessing</i> Teks Ulasan                        | 65 |
| 4.1.6   | Hasil Pembentukan Label Sentimen berbasis Proxy Rating        | 65 |
| 4.1.7   | Hasil Pembagian Dataset                                       | 67 |
| 4.1.8   | Ringkasan Dataset Akhir                                       | 68 |

|        |                                                                 |     |
|--------|-----------------------------------------------------------------|-----|
| 4.2    | Hasil Implementasi dan Evaluasi Model                           | 69  |
| 4.2.1  | Skenario Evaluasi Model                                         | 70  |
| 4.2.2  | Konfigurasi Eksperimen                                          | 70  |
| 4.2.3  | Normalisasi Label dan Fungsi Evaluasi                           | 72  |
| 4.2.4  | Evaluasi GPT <i>prompt-based zero-shot</i>                      | 72  |
| 4.2.5  | Evaluasi Model <i>Zero-Shot</i>                                 | 74  |
| 4.2.6  | Evaluasi Model <i>Fine-Tuned</i>                                | 77  |
| 4.2.7  | Perbandingan Hasil Evaluasi Utama                               | 80  |
| 4.2.8  | Visualisasi Perbandingan Model                                  | 81  |
| 4.2.9  | Analisis Performa per Kelas Sentimen                            | 82  |
| 4.2.10 | Perbandingan Pendekatan <i>Zero-Shot</i> dan <i>Fine-Tuning</i> | 83  |
| 4.2.11 | Evaluasi Per Bahasa                                             | 83  |
| 4.2.12 | Penentuan Model Terbaik                                         | 85  |
| 4.3    | Hasil Pelabelan Dataset Akhir                                   | 86  |
| 4.3.1  | Model Pelabelan yang Digunakan                                  | 86  |
| 4.3.2  | Distribusi Label Sentimen Hasil Pelabelan                       | 87  |
| 4.3.3  | Analisis Kesesuaian Label Model terhadap Proxy Label            | 87  |
| 4.3.4  | Analisis Skor Kepercayaan Model                                 | 89  |
| 4.3.5  | Contoh Kasus Ketidakesuaian Label                               | 90  |
| 4.3.6  | Struktur dan Metadata Dataset Akhir                             | 90  |
| 4.4    | Diskusi dan Pembahasan                                          | 92  |
| 4.4.1  | Penilaian Model Berdasarkan Kriteria Sukses                     | 92  |
| 4.4.2  | Analisis Model Terbaik                                          | 94  |
| 4.4.3  | Perbandingan XLM-RoBERTa dan mBERT                              | 95  |
| 4.4.4  | Analisis GPT <i>prompt-based zero-shot</i>                      | 96  |
| 4.4.5  | Analisis Kelas Netral                                           | 96  |
| 4.4.6  | Analisis Performa Pelabelan dengan Beragam Bahasa               | 97  |
| 4.4.7  | Keterbatasan Label Berbasis Rating                              | 98  |
| 4.4.8  | Keterbatasan Penelitian                                         | 98  |
| 4.4.9  | Ringkasan Pembahasan                                            | 99  |
| 4.5    | Penutup                                                         | 100 |
| BAB V  | PENUTUP                                                         | 101 |
| 5.1    | Kesimpulan                                                      | 101 |
| 5.2    | Saran                                                           | 102 |
|        | DAFTAR PUSTAKA                                                  | 103 |
|        | LAMPIRAN                                                        | 113 |

|                                                                  |            |
|------------------------------------------------------------------|------------|
| <b>Lampiran A: Dataset Ulasan <i>Attraction</i> di Indonesia</b> | <b>113</b> |
| <b>Lampiran B: Repository dan Berkas Pendukung</b>               | <b>115</b> |

## DAFTAR TABEL

|                                                                                                                  |    |
|------------------------------------------------------------------------------------------------------------------|----|
| Tabel 2. 1 Rangkuman Penelitian Terkait Tema Pemanfaatan NLP untuk Labeling Otomatis Ulasan Wisata di Indonesia. | 10 |
| Tabel 2. 2 Perbandingan BERT dan Transformer.                                                                    | 25 |
| Tabel 2. 3 Skema Pemetaan Rating ke Kelas Sentimen.                                                              | 28 |
| Tabel 3. 1 Pemetaan Rating ke Proxy Label Sentimen.                                                              | 47 |
| Tabel 3. 2 Rancangan Pembagian Dataset Penelitian.                                                               | 49 |
| Tabel 3. 3 Model Klasifikasi Sentimen yang Digunakan dalam Penelitian.                                           | 49 |
| Tabel 3. 4 Konfigurasi Hyperparameter pada Proses Fine-Tuning.                                                   | 51 |
| Tabel 3. 5 Instrumen Penelitian yang digunakan untuk proses Penyusunan Tugas Akhir.                              | 55 |
| Tabel 4. 1 Contoh Data Referensi Destinasi Wisata dari ITD Kaggle (Prabowo, 2021) dan DTW Sisparnas 2023.        | 59 |
| Tabel 4. 2 Contoh Temuan dan Pembaruan Koordinat Hasil Cross-Check ITD Kaggle 2021 dan DTW Sisparnas 2023.       | 59 |
| Tabel 4. 3 Cuplikan Raw Data Hasil Web Scraping Tripadvisor.                                                     | 60 |
| Tabel 4. 4 Cuplikan Raw Data Hasil Web Scraping Google Review.                                                   | 61 |
| Tabel 4. 5 Cuplikan Dataset Awal Hasil Penggabungan TripAdvisor dan Google Review.                               | 61 |
| Tabel 4. 6 Deskripsi Kolom Dataset Awal Hasil Akuisisi.                                                          | 62 |
| Tabel 4. 7 Rincian Pengurangan Data pada Tahap Pembersihan Awal.                                                 | 63 |
| Tabel 4. 8 Rekapitulasi Jumlah Data pada Setiap Tahap Penyusunan Dataset.                                        | 63 |
| Tabel 4. 9 Distribusi Bahasa pada Dataset Final.                                                                 | 64 |
| Tabel 4. 10 Perbandingan Label Bahasa Hasil Scraping dan Hasil Deteksi Otomatis.                                 | 64 |
| Tabel 4. 11 Contoh Transformasi Teks Ulasan Sebelum dan Sesudah Preprocessing.                                   | 65 |
| Tabel 4. 12 Distribusi Label Sentimen pada Dataset Final.                                                        | 66 |
| Tabel 4. 13 Distribusi Label Sentimen pada Setiap Bahasa.                                                        | 66 |
| Tabel 4. 14 Pembagian Dataset Final ke dalam Tiga Subset.                                                        | 68 |
| Tabel 4. 15 Distribusi Label Sentimen pada Setiap Subset.                                                        | 68 |
| Tabel 4. 16 Distribusi Bahasa pada Setiap Subset.                                                                | 68 |
| Tabel 4. 17 Struktur Kolom Dataset Final.                                                                        | 69 |
| Tabel 4. 18 Skenario Evaluasi Model.                                                                             | 70 |
| Tabel 4. 19 Konfigurasi Eksperimen Fine-Tuning.                                                                  | 72 |
| Tabel 4. 20 Hasil Evaluasi GPT prompt-based zero-shot.                                                           | 74 |
| Tabel 4. 21 Classification Report GPT prompt-based zero-shot.                                                    | 74 |
| Tabel 4. 22 Hasil Evaluasi mBERT-XNLI Zero-Shot.                                                                 | 76 |
| Tabel 4. 23 Classification Report mBERT-XNLI Zero-Shot.                                                          | 76 |
| Tabel 4. 24 Hasil Evaluasi XLM-RoBERTa Zero-Shot.                                                                | 76 |
| Tabel 4. 25 Classification Report XLM-RoBERTa Zero-Shot.                                                         | 76 |
| Tabel 4. 26 Hasil Evaluasi mBERT Fine-Tuned.                                                                     | 79 |
| Tabel 4. 27 Classification Report mBERT Fine-Tuned.                                                              | 79 |
| Tabel 4. 28 Hasil Evaluasi XLM-RoBERTa Fine-Tuned.                                                               | 79 |
| Tabel 4. 29 Classification Report XLM-RoBERTa Fine-Tuned.                                                        | 79 |
| Tabel 4. 30 Perbandingan Hasil Evaluasi Utama Semua Model.                                                       | 80 |

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| Tabel 4. 31 Perbandingan F1-score Setiap Model pada Masing-masing Kelas Sentimen. | 82 |
| Tabel 4. 32 Perbandingan Macro F1-score Pendekatan Zero-Shot dan Fine-Tuning.     | 83 |
| Tabel 4. 33 Ringkasan Macro F1-score Setiap Model pada Masing-masing Bahasa.      | 83 |
| Tabel 4. 34 Ringkasan Evaluasi Per Bahasa.                                        | 84 |
| Tabel 4. 35 Perbandingan Distribusi Proxy Label dan Label Hasil Model.            | 87 |
| Tabel 4. 36 Distribusi Label Sentimen Hasil Model pada Setiap Bahasa.             | 87 |
| Tabel 4. 37 Matriks Kesesuaian Proxy Label dan Label Hasil Model.                 | 87 |
| Tabel 4. 38 Statistik Skor Kepercayaan Model pada Setiap Kelas Sentimen.          | 89 |
| Tabel 4. 39 Skor Kepercayaan Berdasarkan Kesesuaian dengan Proxy Label.           | 90 |
| Tabel 4. 40 Contoh Ketidakesuaian antara Proxy Label dan Label Hasil Model.       | 90 |
| Tabel 4. 41 Struktur Kolom Dataset Akhir Siap Publikasi.                          | 91 |
| Tabel 4. 42 Metadata Dataset Akhir.                                               | 91 |
| Tabel 4. 43 Peran Setiap Metrik Evaluasi dalam Penilaian Kriteria Sukses.         | 92 |
| Tabel 4. 44 Penilaian Kelima Model terhadap Kriteria Sukses.                      | 93 |

## DAFTAR GAMBAR

|                                                                                                              |    |
|--------------------------------------------------------------------------------------------------------------|----|
| Gambar 2. 1 Alur Web Scraping dengan menggunakan metode API [34].                                            | 13 |
| Gambar 2. 2 Arsitektur Transformer [68].                                                                     | 20 |
| Gambar 2. 3 Fungsionalitas Pre-training pada Arsitektur BERT [68].                                           | 24 |
| Gambar 3. 1 Kerangka Penelitian Pelabelan Sentimen Ulasan Attraction Multibahasa Menggunakan Pendekatan NLP. | 38 |
| Gambar 3. 2 Flowchart Proses Penyusunan Dataset.                                                             | 39 |
| Gambar 3. 3 Hasil visualisasi EDA dari distribusi per-wilayah.                                               | 42 |
| Gambar 3. 4 Diagram Tahapan Pre-Processing Teks.                                                             | 47 |
| Gambar 3. 5 Struktur Data Berbasis Berkas (File-Based) Dataset Ulasan Wisata Bali.                           | 54 |
| Gambar 4. 1 Perbandingan Distribusi Bahasa Sebelum dan Sesudah Deteksi Otomatis.                             | 65 |
| Gambar 4. 2 Distribusi Label Sentimen pada Setiap Bahasa.                                                    | 67 |
| Gambar 4. 3 Verifikasi Konsistensi Proporsi Sentimen dan Bahasa Antar Subset.                                | 68 |
| Gambar 4. 4 Diagram Alur Konfigurasi Eksperimen dan Penggunaan Data oleh Lima Model.                         | 71 |
| Gambar 4. 5 Diagram Alur Evaluasi GPT prompt-based zero-shot.                                                | 73 |
| Gambar 4. 6 Diagram Alur Evaluasi Model Zero-Shot Berbasis NLI (mBERT-XNLI dan XLM-RoBERTa).                 | 75 |
| Gambar 4. 7 Diagram Alur Pelatihan dan Evaluasi Model Fine-Tuned (mBERT dan XLM-RoBERTa).                    | 78 |
| Gambar 4. 8 Perbandingan Accuracy Semua Model.                                                               | 81 |
| Gambar 4. 9 Perbandingan Macro F1-score Semua Model.                                                         | 81 |
| Gambar 4. 10 Perbandingan F1-score per Kelas Sentimen.                                                       | 82 |
| Gambar 4. 11 Perbandingan Macro F1-score pada setiap Model Bahasa.                                           | 85 |
| Gambar 4. 12 Kesesuaian Proxy Label dan Label Model pada Keseluruhan Dataset.                                | 88 |
| Gambar 4. 13 Distribusi Label Sentimen dan Skor Kepercayaan pada Dataset Akhir.                              | 89 |

## DAFTAR SINGKATAN

|         |                                                                  |
|---------|------------------------------------------------------------------|
| NLP     | : <i>Natural Language Processing</i>                             |
| BERT    | : <i>Bidirectional Encoder Representations from Transformers</i> |
| GPT     | : <i>Generative Pre-trained Transformer</i>                      |
| mBERT   | : <i>multilingual BERT</i>                                       |
| XNLI    | : <i>Cross-lingual Natural Language Inference</i>                |
| RoBERTa | : <i>Robustly Optimized BERT Pretraining Approach</i>            |
| ABSA    | : <i>Aspect-Based Sentiment Analysis</i>                         |
| API     | : <i>Application Programming Interface</i>                       |
| BoW     | : <i>Bag-of-Words</i>                                            |
| CSV     | : <i>Comma Separated Values</i>                                  |
| DPN     | : <i>Destinasi Pariwisata Nasional</i>                           |
| ITD     | : <i>Indonesia Tourism Destination</i>                           |
| DTW     | : <i>Daya Tarik Wisata</i>                                       |
| HTML    | : <i>HyperText Markup Language</i>                               |
| EDA     | : <i>Exploratory Data Analysis</i>                               |
| GPU     | : <i>Graphics Processing Unit</i>                                |
| GRU     | : <i>Gated Recurrent Unit</i>                                    |
| HTTP    | : <i>Hypertext Transfer Protocol</i>                             |
| TIK     | : <i>Teknologi Informasi dan Komunikasi</i>                      |
| JSON    | : <i>JavaScript Object Notation</i>                              |
| LLM     | : <i>Large Language Model</i>                                    |
| LSTM    | : <i>Long Short-Term Memory</i>                                  |
| MLM     | : <i>Masked Language Modeling</i>                                |
| MNLI    | : <i>Multi-Genre Natural Language Inference</i>                  |
| GLUE    | : <i>General Language Understanding Evaluation</i>               |
| NER     | : <i>Named Entity Recognition</i>                                |
| NLI     | : <i>Natural Language Inference</i>                              |
| NLG     | : <i>Natural Language Generation</i>                             |
| NSP     | : <i>Next Sentence Prediction</i>                                |
| NLU     | : <i>Natural Language Understanding</i>                          |
| PDB     | : <i>Produk Domestik Bruto</i>                                   |
| POA     | : <i>Part of Article</i>                                         |
| QA      | : <i>Question-Answering</i>                                      |
| RNN     | : <i>Recurrent Neural Network</i>                                |
| SQuAD   | : <i>Stanford Question Answering Dataset</i>                     |
| SVM     | : <i>Support Vector Machine</i>                                  |
| TF-IDF  | : <i>Term Frequency-Inverse Document Frequency</i>               |
| URL     | : <i>Uniform Resource Locator</i>                                |
| XLSX    | : <i>Microsoft Excel Open XML Spreadsheet</i>                    |
| XLM-R   | : <i>XLM-RoBERTa</i>                                             |
| XML     | : <i>Extensible Markup Language</i>                              |

## DAFTAR ISTILAH

|                         |                                                                                                                                                                                                              |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Actor</i>            | Skrip siap pakai dalam platform seperti Apify yang digunakan untuk melakukan <i>web scraping</i> otomatis.                                                                                                   |
| <i>Amenitas</i>         | Fasilitas pendukung yang tersedia di destinasi wisata.                                                                                                                                                       |
| Anotasi Manual          | Proses pelabelan data secara manual oleh manusia untuk membuat dataset berlabel.                                                                                                                             |
| <i>Attraction</i>       | Merujuk pada atraksi wisata atau tempat-tempat menarik bagi wisatawan.                                                                                                                                       |
| <i>Batch Size</i>       | Jumlah sampel pelatihan yang diproses dalam satu iterasi selama pelatihan model.                                                                                                                             |
| <i>Bidirectional</i>    | Kemampuan model (seperti BERT) untuk memproses <i>input</i> teks dari dua arah (kiri ke kanan dan kanan ke kiri) secara bersamaan untuk memahami konteks penuh dari sebuah kata.                             |
| <i>Bot/Web Crawler</i>  | Program otomatis yang digunakan untuk menjelajahi situs web dan mengambil informasi.                                                                                                                         |
| <i>Cloud-based</i>      | Sistem atau platform yang beroperasi dan diakses melalui internet, tidak disimpan secara lokal.                                                                                                              |
| <i>Confidence Score</i> | Nilai probabilitas yang dihasilkan oleh model, menunjukkan seberapa besar kemungkinan suatu prediksi akurat.                                                                                                 |
| <i>Confusion Matrix</i> | Tabel yang membandingkan hasil prediksi model dengan label aktual pada setiap kelas, digunakan untuk menghitung metrik evaluasi seperti <i>accuracy</i> , <i>precision</i> , <i>recall</i> , dan <i>F1</i> . |
| <i>Context-aware</i>    | Kemampuan model untuk memahami makna kata berdasarkan lingkungan atau konteks kalimatnya.                                                                                                                    |
| <i>Decoder</i>          | Bagian dari arsitektur <i>Transformer</i> yang bekerja pada urutan output yang diinginkan selama pelatihan dan memprediksi item berikutnya dalam urutan.                                                     |
| <i>Deep Learning</i>    | Cabang dari <i>machine learning</i> yang menggunakan jaringan saraf tiruan dengan banyak lapisan ( <i>deep neural networks</i> ) untuk mempelajari representasi data secara hierarkis.                       |
| Digital Era             | Periode saat ini yang ditandai oleh dominasi teknologi digital dalam berbagai aspek kehidupan.                                                                                                               |

|                            |                                                                                                                                                                                                                                  |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Distant Supervision</i> | Teknik pelabelan otomatis yang memanfaatkan sumber sinyal yang sudah tersedia (misalnya rating) sebagai perkiraan label, alih-alih anotasi manual oleh manusia. Label yang dihasilkan disebut <i>distant supervision label</i> . |
| <i>Domain-specific</i>     | Model atau pendekatan yang disesuaikan atau dilatih khusus untuk domain atau area tertentu (misalnya, pariwisata).                                                                                                               |
| <i>E-commerce</i>          | Aktivitas jual beli produk atau layanan melalui internet.                                                                                                                                                                        |
| <i>E-tourism</i>           | Proses digitalisasi seluruh rantai nilai dalam industri perjalanan dan pariwisata untuk meningkatkan efisiensi dan efektivitas operasional melalui TIK.                                                                          |
| <i>Edge Devices</i>        | Perangkat komputasi yang beroperasi di "tepi" jaringan, dekat dengan sumber data, seperti perangkat seluler atau sensor.                                                                                                         |
| <i>Encoder</i>             | Bagian dari arsitektur <i>Transformer</i> yang bekerja terutama pada urutan <i>input</i> .                                                                                                                                       |
| <i>Entailment</i>          | Hubungan logis di mana satu kalimat (hipotesis) dapat disimpulkan dari kalimat lain (premis). Digunakan dalam NLI.                                                                                                               |
| <i>Epoch</i>               | Satu siklus penuh di mana seluruh dataset pelatihan dilewatkan melalui model.                                                                                                                                                    |
| <i>Fair Use</i>            | Doktrin hukum yang memungkinkan penggunaan bahan berhak cipta dalam keadaan tertentu tanpa izin dari pemegang hak cipta, seringkali untuk tujuan pendidikan atau riset.                                                          |
| <i>Few-shot Learning</i>   | Pendekatan pembelajaran mesin di mana model dapat belajar dan membuat prediksi yang baik dari sejumlah kecil contoh data berlabel.                                                                                               |
| <i>Fine-tuning</i>         | Proses melatih ulang model <i>pre-trained</i> pada dataset spesifik suatu tugas agar bobot model menyesuaikan diri dengan karakteristik data dan tujuan tugas tersebut.                                                          |
| <i>Framework</i>           | Lingkungan perangkat lunak yang menyediakan struktur dasar untuk membangun aplikasi.                                                                                                                                             |
| Generalisasi               | Kemampuan model pembelajaran mesin untuk berkinerja baik pada data baru yang belum pernah dilihat sebelumnya, setelah dilatih pada dataset tertentu.                                                                             |
| <i>Ground Truth</i>        | Label acuan yang dianggap benar, umumnya diperoleh melalui anotasi manual, yang digunakan sebagai pembanding untuk mengevaluasi ketepatan prediksi model.                                                                        |

|                             |                                                                                                                                                                              |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Google Colab                | Platform <i>cloud</i> berbasis browser yang mendukung GPU dan integrasi dengan Google Drive, sering digunakan untuk eksperimen <i>machine learning</i> .                     |
| Hugging Face                | Pustaka <i>open-source</i> berbasis Python yang menyediakan antarmuka tinggi untuk bekerja dengan berbagai model <i>pre-trained</i> berbasis arsitektur <i>Transformer</i> . |
| <i>Hypothesis Statement</i> | Kalimat atau frasa yang digunakan dalam <i>zero-shot classification</i> untuk menguji hubungan semantik dengan teks input.                                                   |
| IndoBERT                    | Model pralatih berbasis arsitektur BERT yang dikembangkan khusus untuk Bahasa Indonesia.                                                                                     |
| Interpretabilitas           | Sejauh mana keputusan atau prediksi model <i>machine learning</i> dapat dipahami atau dijelaskan oleh manusia.                                                               |
| Kaggle                      | Platform online untuk ilmu data dan pembelajaran mesin yang menyediakan dataset publik.                                                                                      |
| Klasifikasi Sentimen        | Tugas dalam NLP untuk mengkategorikan teks ke dalam sentimen positif, negatif, atau netral.                                                                                  |
| Kontekstual                 | Berkaitan dengan atau tergantung pada konteks di mana sesuatu terjadi atau muncul.                                                                                           |
| Korpus                      | Kumpulan besar teks atau data bahasa yang digunakan untuk melatih model bahasa.                                                                                              |
| <i>Labeling</i> (Pelabelan) | Proses penandaan data tidak terstruktur (seperti teks ulasan) dengan informasi tambahan, seperti kategori atau nilai sentimen.                                               |
| <i>Layer Freezing</i>       | Teknik dalam <i>fine-tuning</i> di mana beberapa lapisan bawah model tetap dibekukan (tidak dilatih ulang) untuk mempertahankan pemahaman bahasa umum.                       |
| <i>Learning Rate</i>        | Ukuran langkah yang digunakan dalam mengoptimalkan model selama pelatihan.                                                                                                   |
| <i>Low Resource</i>         | Merujuk pada bahasa atau domain dengan ketersediaan data berlabel yang terbatas.                                                                                             |
| <i>Machine Learning</i>     | Cabang dari kecerdasan buatan yang memungkinkan sistem belajar dari data dan membuat prediksi atau keputusan tanpa pemrograman eksplisit.                                    |
| <i>Multi-head Attention</i> | Pengembangan dari mekanisme self-attention yang memungkinkan model memperhatikan berbagai representasi semantik secara paralel.                                              |

|                             |                                                                                                                                                                                                                                                                                                                      |
|-----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Multilingual</i>         | Mampu menangani atau memproses lebih dari satu bahasa.                                                                                                                                                                                                                                                               |
| <i>Noise</i>                | Data yang tidak relevan, salah, atau mengganggu dalam dataset.                                                                                                                                                                                                                                                       |
| <i>Non-sekuensial</i>       | Tidak berurutan; dalam konteks <i>Transformer</i> , berarti dapat memproses <i>input</i> secara paralel.                                                                                                                                                                                                             |
| <i>Overfitting</i>          | Kondisi di mana model <i>machine learning</i> berkinerja sangat baik pada data pelatihan tetapi buruk pada data baru yang tidak terlihat karena terlalu banyak belajar dari " <i>noise</i> " dalam data pelatihan.                                                                                                   |
| <i>Pipeline</i>             | Urutan proses atau tahapan yang saling berhubungan untuk mencapai suatu hasil.                                                                                                                                                                                                                                       |
| <i>Pre-trained Model</i>    | Model pembelajaran mesin yang telah dilatih pada dataset yang sangat besar dan kemudian dapat disesuaikan untuk tugas-tugas spesifik.                                                                                                                                                                                |
| Probabilitas                | Ukuran kemungkinan terjadinya suatu peristiwa.                                                                                                                                                                                                                                                                       |
| <i>Prompt-based</i>         | Pendekatan yang mengarahkan model bahasa besar ( <i>Large Language Model</i> ) melalui instruksi teks ( <i>prompt</i> ) untuk menghasilkan prediksi, tanpa melakukan pelatihan atau pembaruan bobot model.                                                                                                           |
| <i>Proxy Label</i>          | Label sentimen sementara yang diturunkan dari sinyal tidak langsung, yaitu rating ulasan (1–2 = negatif, 3 = netral, 4–5 = positif), dan disimpan pada kolom <i>proxy_sentiment</i> . Label ini berperan sebagai <i>distant supervision label</i> , bukan hasil anotasi manual manusia maupun <i>gold standard</i> . |
| <i>Real-time</i>            | Pemrosesan atau respons yang terjadi segera setelah input diterima.                                                                                                                                                                                                                                                  |
| Redundansi Data             | Duplikasi data yang tidak perlu dalam suatu sistem.                                                                                                                                                                                                                                                                  |
| <i>Rule-based</i>           | Pendekatan yang mengandalkan seperangkat aturan yang telah ditentukan sebelumnya.                                                                                                                                                                                                                                    |
| <i>Self-attention</i>       | Mekanisme dalam arsitektur <i>Transformer</i> yang memungkinkan model mempertimbangkan semua posisi dalam input secara bersamaan dan menghitung bobot relevansi antara setiap pasangan kata.                                                                                                                         |
| <i>Semi-structured data</i> | Data yang tidak sesuai dengan model data relasional tabel tetapi memiliki tag atau penanda lain yang mengorganisasikan elemen-elemen di dalamnya.                                                                                                                                                                    |
| <i>Sentiment Analysis</i>   | Proses penggunaan NLP untuk menentukan nada emosional (positif, negatif, netral) dari teks.                                                                                                                                                                                                                          |

|                            |                                                                                                                                                                                                                                                                  |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Slug</i>                | Bagian dari URL yang mengidentifikasi sumber daya secara unik dan mudah dibaca oleh manusia.                                                                                                                                                                     |
| <i>Spreadsheet</i>         | Dokumen yang menyimpan data dalam baris dan kolom, seperti Microsoft Excel atau Google Sheets.                                                                                                                                                                   |
| <i>State-of-the-art</i>    | Mengacu pada metode, teknik, atau hasil terbaru dan paling canggih dalam suatu bidang.                                                                                                                                                                           |
| <i>Stratified Sampling</i> | Teknik pembagian data yang mempertahankan proporsi kelas pada setiap subset ( <i>train</i> , <i>validation</i> , dan <i>test</i> ) agar distribusinya tetap seimbang, pada penelitian ini stratifikasi dilakukan berdasarkan bahasa dan <i>proxy_sentiment</i> . |
| <i>Supervised Learning</i> | Jenis pembelajaran mesin di mana model dilatih pada dataset yang telah diberi label ( <i>input</i> dipasangkan dengan <i>output</i> yang benar).                                                                                                                 |
| <i>Transfer Learning</i>   | Teknik <i>machine learning</i> di mana model yang dilatih untuk satu tugas digunakan sebagai titik awal untuk tugas lain.                                                                                                                                        |
| <i>Unsupervised</i>        | Jenis pembelajaran mesin di mana model belajar dari data yang tidak berlabel, mencari pola atau struktur dalam data itu sendiri.                                                                                                                                 |
| <i>Tokenizer</i>           | Komponen yang memecah teks menjadi unit-unit yang lebih kecil (token) yang dapat diproses oleh model NLP.                                                                                                                                                        |
| <i>Web Scraping</i>        | Metode pengambilan data secara otomatis dari situs web.                                                                                                                                                                                                          |
| <i>Zero-shot Learning</i>  | Pendekatan di mana model melakukan prediksi terhadap kelas atau tugas tanpa pernah dilatih menggunakan contoh berlabel spesifik untuk kelas tersebut, dengan mengandalkan pengetahuan yang sudah dipelajari saat <i>pre-training</i> .                           |

## DAFTAR LAMPIRAN

|                                                           |     |
|-----------------------------------------------------------|-----|
| Lampiran A: Dataset Ulasan <i>Attraction</i> di Indonesia | 113 |
| Lampiran B: Repository dan Berkas Pendukung               | 115 |