

DAFTAR ISI

SURAT PERNYATAAN	iii
KARYA ASLI TUGAS AKHIR	iii
PERNYATAAN BEBAS PLAGIASI	iv
BAB I	1
PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Tahapan Penelitian	4
1.7 Sistematika Penulisan	6
BAB II	8
TINJAUAN LITERATUR	8
2.1 Analisis Sentimen	8
2.2 Ulasan Google Maps.....	8
2.3 <i>Bidirectional Encoder Representations from Transformers (BERT)</i>	9
2.4 <i>Fine-tuning BERT</i>	11
2.5 <i>IndoBERT</i>	12
2.6 <i>Data Collecting</i>	12
2.7 <i>Apify</i> 13	
2.7.1 <i>Actor Google Maps Review Scraper</i>	13
2.7.2 <i>Lexicon Indonesian Sentiment Lexicon (INSET)</i>	14
2.8 <i>Data Labeling</i>	14
2.8.1 <i>N-Gram</i>	15
2.9 <i>Data Pre-processing</i>	15
2.9.1 <i>Cleansing</i>	16
2.9.2 <i>Casefolding</i>	16
2.9.3 <i>Normalization</i>	16
2.9.4 <i>Tokenization</i>	16
2.9.5 <i>Stopword Removal</i>	16
2.9.6 <i>Stemming</i>	17
2.10 <i>Data Resampling</i>	17
2.11 <i>Data Splitting</i>	17
2.12 <i>Confusion Matrix</i>	18
2.13 <i>Cross-Industry Standard Process for Data Mining (CRISP – DM)</i>	19

2.14	<i>Python</i>	19
2.15	<i>Google Colaboratory</i>	19
2.16	Penelitian Terdahulu	20
BAB III		21
METODOLOGI PENELITIAN		21
3.1	Tahapan Penelitian	21
3.2	Identifikasi Masalah	21
3.3	Studi Literatur	22
3.4	Pengumpulan Data	22
3.5	<i>Pre-processing Data</i>	22
3.5.1	<i>Cleansing</i>	23
3.5.2	<i>Casefolding</i>	24
3.5.3	<i>Normalization</i>	25
3.5.4	<i>Tokenizing</i>	26
3.5.5	<i>Stopword Removal</i>	26
3.5.6	<i>Stemming</i>	27
3.6	<i>Data Labeling</i>	29
3.6.1	<i>Indonesian Sentiment Lexicon Dengan Dukungan N-Gram</i>	30
3.6.2	<i>Indonesian Sentiment Lexicon Tanpa Dukungan N-Gram</i>	34
3.6.3	<i>Berdasarkan Rating</i>	37
3.7	<i>Data Resampling</i>	37
3.7.1	<i>Resampling Data Lexicon INSET dengan Dukungan N-Gram</i>	38
3.7.2	<i>Resampling Data Lexicon INSET Tanpa Dukungan N-Gram</i>	38
3.8	<i>Data Splitting</i>	39
3.9	Tokenisasi IndoBERT	40
3.9.1	<i>Wordpiece Tokenizer</i>	41
3.9.2	Penambahan Token Spesial	42
3.9.3	<i>Encoding Token ID</i>	42
3.9.4	<i>Encoding Token Embedding</i>	44
3.9.5	<i>Encoding Input Embedding</i>	44
3.9.6	<i>Encoder Layer</i>	47
3.9.7	<i>Linear Layer</i>	53
3.9.8	<i>Softmax</i>	55
3.10	<i>Fine-tuning Hyperparameter</i>	55
3.10.1	<i>Learning Rate</i>	56
3.10.2	<i>Batch Size</i>	56
3.10.3	<i>Epoch</i>	57

3.11	Evaluasi Model	57
3.11.1	Akurasi	57
3.11.2	Presisi	58
3.11.3	<i>Recall</i>	59
3.11.4	<i>F1-Score</i>	60
3.12	Evaluasi Optimasi	61
3.12.1	Sebelum Optimasi	62
3.12.2	Sesudah Optimasi	62
3.13	<i>Deploy Model</i>	63
BAB IV		64
HASIL PENGUJIAN DAN PEMBAHASAN		64
4.1	Implementasi.....	64
4.2	Pengumpulan Data	64
4.2.1	<i>Cleansing</i>	65
4.2.2	<i>Casefolding</i>	65
4.2.3	<i>Normalization</i>	66
4.2.4	<i>Tokenizing</i>	66
4.2.5	<i>Stopword Removal</i>	67
4.2.6	<i>Stemming</i>	68
4.2.7	<i>Data Labeling Untuk Lexicon INSET dengan Dukungan N-Gram</i>	69
4.2.8	<i>Data Labeling Untuk Lexicon INSET Tanpa Dukungan N-Gram</i>	70
4.2.9	<i>Data Labeling Berdasarkan Rating</i>	71
4.2.10	<i>Data Resampling</i>	72
4.2.11	<i>Data Splitting</i>	72
4.2.12	Tokenisasi IndoBERT	73
4.2.13	Evaluasi Pretrained IndoBERT-Base-P1	74
4.2.14	<i>Fine-tuning dan Evaluasi Hyperparameter</i>	75
4.2.15	<i>Deploy</i>	77
4.3	Hasil 80	
4.3.1	<i>Data Labeling Lexicon INSET dengan Dukungan N-Gram</i>	80
4.3.2	<i>Data Labeling Untuk Lexicon INSET Tanpa Dukungan N-Gram</i>	81
4.3.3	Perbandingan <i>Data Labeling Lexicon INSET dengan Dukungan N-Gram vs. Data Labeling Untuk Lexicon INSET Tanpa Dukungan N-Gram</i>	81
4.3.4	<i>Data Labeling Berdasarkan Rating</i>	83
4.3.5	Evaluasi Parameter Rekomendasi Devlin et al. (2018) Menggunakan <i>Lexicon INSET dengan Dukungan N-Gram Sebelum Upsampling</i>	83

4.3.6	Evaluasi Parameter Rekomendasi Devlin et al. (2018) Menggunakan <i>Lexicon</i> INSET Tanpa Dukungan <i>N-Gram</i> Sebelum <i>Upsampling</i>	84
4.3.7	<i>Data Resampling Lexicon INSET</i> dengan Dukungan <i>N-Gram</i>	85
4.3.8	<i>Data Resampling Lexicon INSET</i> Tanpa Dukungan <i>N-Gram</i>	86
4.3.9	<i>Data Splitting</i>	87
4.3.10	Evaluasi <i>Pre-trained</i> Model IndoBERT-base-p1 Menggunakan <i>Lexicon</i> INSET dengan Dukungan <i>N-Gram</i>	87
4.3.11	Evaluasi <i>Pre-trained</i> Model IndoBERT-base-p1 Menggunakan <i>Lexicon</i> INSET Tanpa Dukungan <i>N-Gram</i>	89
4.3.12	Evaluasi Parameter Rekomendasi Devlin et al. (2018) Menggunakan <i>Lexicon</i> INSET dengan Dukungan <i>N-Gram</i>	90
4.3.13	Evaluasi Parameter Rekomendasi Devlin et al. (2018) Menggunakan <i>Lexicon</i> INSET Tanpa Dukungan <i>N-Gram</i>	91
4.3.14	Evaluasi Parameter Rekomendasi Devlin et al. (2018) Menggunakan <i>Data Labeling</i> Berdasarkan <i>Rating</i>	91
4.3.15	Perbandingan Evaluasi Sebelum <i>Upsampling</i> vs. Sesudah <i>Upsampling</i> 92	
4.3.16	Perbandingan Evaluasi Optimasi	93
4.3.17	<i>Implementasi User Interface</i>	94
4.4	Pembahasan	96
4.4.1	Perbandingan <i>Labeling Lexicon-based</i> vs. Berdasarkan <i>Rating</i>	96
4.4.2	Pengaruh Penggunaan <i>N-Gram</i>	97
4.4.3	Perbandingan Sebelum Optimasi (<i>Pre-trained</i>) vs. Optimasi (<i>Fine-tuning</i>)	97
4.4.4	Konfigurasi <i>Hyperparameter</i> Terbaik	98
4.4.5	Korelasi Antara Parameter dengan <i>Matrix</i>	98
BAB V		100
PENUTUP		100
5.1	Kesimpulan	100
5.2	Saran	100
DAFTAR PUSTAKA		101

DAFTAR TABEL

Tabel 2. 1 Confusion Matrix (Lastetria., 2024).....	18
Tabel 2. 2 Penelitian Terdahulu.....	20
Tabel 3. 1 Contoh Hasil Pengumpulan Data	22
Tabel 3. 2 Contoh Penggunaan Cleansing	24
Tabel 3. 3 Contoh Penggunaan Casefolding	24
Tabel 3. 4 Contoh Penggunaan Normalization	25
Tabel 3. 5 Contoh Penggunaan Tokenizing.....	26
Tabel 3. 6 Contoh Penggunaan Stopword Removal	27
Tabel 3. 7 Contoh Penggunaan Stemming	28
Tabel 3. 8 Contoh entri lexicon INSET.....	29
Tabel 3. 9 Hasil Generate <i>N-Gram</i>	32
Tabel 3. 10 Pencocokan Kamus Lexicon	32
Tabel 3. 11 Hasil Generate <i>N-Gram</i>	35
Tabel 3. 12 Pencocokan Kamus Lexicon	35
Tabel 3. 14 Implementasi Wordpiece Tokenizer	41
Tabel 3. 15 Penambahan Token ID	42
Tabel 3. 16 Penambahan Token Embedding (5 Dimensi Pertama).....	44
Tabel 3. 17 Position Embedding (5 Dimensi Pertama).....	45
Tabel 3. 18 Segment Embedding (5 Dimensi Pertama)	46
Tabel 3. 19 Input Embedding (5 Dimensi Pertama).....	47
Tabel 3. 20 Contoh Hasil Multi-Head Attention Layer ke-0 (5 Dimensi Pertama)	48
Tabel 3. 21 Contoh Hasil Multi-Head Attention Layer ke-0 Yang Dinormalisasi (5 Dimensi Pertama).....	49
Tabel 3. 22 Contoh Hasil Feed Forward Netrok Layer ke-0 (5 Dimensi Pertama)	50
Tabel 3. 23 Contoh Hasil Feed Forward Netrok Layer ke-0 Yang Dinormalisasi (5 Dimensi Pertama).....	50
Tabel 3. 24 Contoh Output Embedding Tiap Layer 0 Sampai Layer 6 (5 Dimensi Pertama)	51
Tabel 3. 25 Lanjutan Contoh Output Embedding Tiap Layer 7 Sampai Layer 12 (5 Dimensi Pertama).....	53
Tabel 3. 26 Input Linear Layer (5 Dimensi Pertama)	54
Tabel 3. 27 Simulasi Presisi	58
Tabel 3. 28 Ilustrasi Prediksi Model	59
Tabel 3. 29 Rencana Perbandingan Evaluasi Optimal	62
Tabel 4. 1 Contoh Hasil Pengumpulan Data	64
Tabel 4. 2 Contoh Data Positive Dengan Pelabelan Lexicon Didukung N-Gram	80
Tabel 4. 3 Contoh Data Neutral Dengan Pelabelan Lexicon Didukung N-Gram	80
Tabel 4. 4 Contoh Data Negative Dengan Pelabelan Lexicon Didukung N-Gram.....	81
Tabel 4. 5 Perbedaan Polaritas Pelabelan Lexicon Dengan N-Gram vs. Tanpa N-Gram ...	82
Tabel 4. 6 Evaluasi Grid Search Data Labeling Dengan Dukungan N-Gram Sebelum Upsampling.....	83

Tabel 4. 7 Evaluasi Grid Search Data Labeling Tanpa Dukungan N-Gram Sebelum Upsampling.....	84
Tabel 4. 8 Evaluasi Pre-trained Model IndoBERT-base-p1 dengan Dukungan N-Gram....	88
Tabel 4. 9 Evaluasi Pre-trained Model IndoBERT-base-p1 Tanpa Dukungan N-Gram	89
Tabel 4. 10 Evaluasi Grid Search Data Labeling N-Gram.....	90
Tabel 4. 11 Evaluasi Grid Search Data Labeling Tanpa N-Gram	91
Tabel 4. 12 Evaluasi Grid Search Data Labeling Berdasarkan Rating	92
Tabel 4. 13 Perbandingan Evaluasi Parameter Terbaik.....	93

DAFTAR GAMBAR

Gambar 2. 1 Arsitektur Transformer (Vaswani et al., 2017)	9
Gambar 2. 2 Representasi Input pada BERT (Devlin et al., 2018)	10
Gambar 2. 3 Proses Mekanisme Pre-training dan Fine-tuning (Devlin et al. 2018).....	11
Gambar 2. 4 Proses Fine-tuning BERT (Devlin et al., 2018)	11
Gambar 2. 5 Model Proses Penambahan Data CRISP-DM (Suhanda et al., 2020).....	19
Gambar 3. 1 Tahapan Penelitian	21
Gambar 3. 2 Diagram Alur Proses Pre-Processing	22
Gambar 3. 3 Diagram Alur Proses Cleansing	23
Gambar 3. 4 Diagram Alur Proses Casefolding	24
Gambar 3. 5 Diagram Alur Proses Normalization	25
Gambar 3. 6 Diagram Alur Proses Tokenizing	26
Gambar 3. 7 Diagram Alur Proses Stopword Removal	27
Gambar 3. 8 Diagram Alur Proses Stemming.....	28
Gambar 3. 9 Diagram Alur Proses Data Labeling Lexicon Dengan <i>N-Gram</i>	30
Gambar 3. 10 Diagram Alur Proses Data Labeling Lexicon Tanpa Dukungan <i>N-Gram</i> ...	34
Gambar 3. 11 Diagram Alur Proses Data Labeling Berdasarkan Rating	37
Gambar 3. 12 Diagram Alur Proses Data Resampling Untuk Lexicon INSET Dengan <i>N-Gram</i>	38
Gambar 3. 13 Diagram Alur Proses Data Resampling Untuk Lexicon INSET Tanpa <i>N-Gram</i>	39
Gambar 3. 14 Diagram Alur Proses Data Splitting.....	40
Gambar 3. 15 Diagram Alur Proses Tokenizing Dengan IndoBERT.....	40
Gambar 3. 16 Contoh Vocab WordPiece Tokenizer	41
Gambar 3. 17 Penambahan Token Khusus	42
Gambar 3. 18 Konversi ke Token ID	43
Gambar 3. 19 Implementasi PADDING dan Attention Mask.....	43
Gambar 3. 20 Diagram Alur Proses Eksplorasi Hyperparameter	56
Gambar 3. 21 Diagram Alur Proses Deploy Model	63
Gambar 3. 22 Wireframe User Interface	63
Gambar 4. 1 Actor Google Maps Reviews Scraper	64
Gambar 4. 2 Sampel Data Hasil Cleansing.....	65
Gambar 4. 3 Hasil Casefolding	66
Gambar 4. 4 Sampel Data Hasil Normalization.....	66
Gambar 4. 5 Sampel Data Hasil Tokenizing	67
Gambar 4. 6 Sampel Data Hasil Stopword Removal.....	68
Gambar 4. 7 Sampel Data Hasil Stemming	68
Gambar 4. 8 Distribusi Data Labeling Lexicon Dengan Dukungan <i>N-Gram</i>	80
Gambar 4. 9 Distribusi Data Labeling Lexicon Tanpa <i>N-Gram</i>	81
Gambar 4. 10 Distribusi Data Labeling Berdasarkan Rating.....	83
Gambar 4. 11 Distribusi Data Untuk Upsampling Dengan Dukungan <i>N-Gram</i>	85
Gambar 4. 12 Hasil Upsampling Dengan Dukungan <i>N-Gram</i>	85

Gambar 4. 13	Distribusi Data Untuk Upsampling Tanpa Dukungan N-Gram	86
Gambar 4. 14	Hasil Upsampling Tanpa Dukungan N-Gram	86
Gambar 4. 15	Distribusi Data Splitting untuk Evaluasi Fine-tuning	87
Gambar 4. 16	<i>Evaluasi Pre-trained Model IndoBERT-base-p1 dengan Dukungan N-Gram</i>	87
Gambar 4. 17	<i>Evaluasi Pre-trained Model IndoBERT-base-p1 Tanpa Dukungan N-Gram</i>	89
Gambar 4. 18	Grafik Perbandingan Sebelum Upsampling vs. Sesudah Upsampling	92
Gambar 4. 19	User Interface Ulasan Satuan	95
Gambar 4. 20	User Interface Ulasan Batch.....	95
Gambar 4. 21	Heatmap Korelasi Antar Parameter dan Matrix	98

DAFTAR MODUL PROGRAM

Modul Program 4. 1	Cleansing.....	65
Modul Program 4. 2	Casefolding	65
Modul Program 4. 3	Normalization	66
Modul Program 4. 4	Tokenizing.....	66
Modul Program 4. 5	Stopword Removal.....	67
Modul Program 4. 6	Stemming	68
Modul Program 4. 7	Fungsi Load Lexicon	69
Modul Program 4. 8	Fungsi Generate <i>N-Gram</i>	69
Modul Program 4. 9	Fungsi Labeling Lexicon	70
Modul Program 4. 10	Fungsi Run Batch Data Labeling dengan <i>N-Gram</i>	70
Modul Program 4. 11	Fungsi Unigram.....	71
Modul Program 4. 12	Fungsi Run Batch Data Labeling Tanpa <i>N-Gram</i>	71
Modul Program 4. 13	Data Labeling Berdasarkan Rating	72
Modul Program 4. 14	Fungsi Penggabungan Data.....	72
Modul Program 4. 15	Fungsi Splitting.....	73
Modul Program 4. 16	Fungsi Tokenizing dengan IndoBERT	73
Modul Program 4. 17	Test Model Pretrained IndoBERT-base-p1	75
Modul Program 4. 18	Fungsi Pelatihan dengan Grid Search	77
Modul Program 4. 19	app.py.....	79