

DAFTAR ISI

HALAMAN PENGESAHAN TESIS	iii
PERNYATAAN.....	iv
PRAKATA.....	v
DAFTAR ISI.....	vii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
INTISARI.....	xii
ABSTRACT.....	xiii
BAB I PENDAHULUAN	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah	6
1.3 Batasan Masalah.....	6
1.4 Tujuan.....	7
1.5 Manfaat.....	7
BAB II TINJAUAN PUSTAKA.....	9
BAB III LANDASAN TEORI.....	14
3.1 Big-Five Personality.....	14
3.2 Natural Language Processing (NLP).....	15
3.3 Pra-proses Data Teks dan Word Embedding	17
3.3.1 Case folding.....	17
3.3.2 Normalisasi teks	17
3.3.3 Penghapusan tanda baca dan karakter khusus	17
3.3.4 Word Embedding.....	18
3.3.5 FastText Word Embedding.....	18
3.4 Data Augmentation	18
3.4.1 Augmentasi non-kontekstual	19
3.4.2 Augmentasi semi-kontekstual	20

3.4.3 Augmentasi kontekstual	21
3.5 Transformer	21
3.5.1 Encoder	22
3.5.2 Decoder	23
3.5.3 BERT	23
3.5.4 IndoBERT	30
3.6 Metode Klasifikasi	30
3.6.1 Multi-label classification	30
3.6.2 Fine-tuning BERT	31
3.6.3 Logistic Regression	33
3.6.4 Support Vector Machine (SVM)	33
3.6.5 Extreme Gradient Boosting (XGBoost)	34
3.7 Evaluasi Model	35
3.7.1 Split dataset untuk pengecekan performa	36
3.7.2 ConfusionMatrix	37
BAB IV METODOLOGI PENELITIAN	40
4.1 Deskripsi Penelitian	40
4.2 Alat & Bahan	41
4.2.1 Alat	41
4.2.2 Bahan	42
4.3 Prosedur Kerja	43
4.4 Pengumpulan Data	44
4.5 Pemberian Label Data Uji	45
4.6 Data Augmentation	46
4.7 Pra-pemrosesan Data	48
4.8 Tokenize	51
4.8.1 Tokenize Word Embedding	51

4.9	Proses Modelling	52
4.9.1	Logistic Regression	54
4.9.2	Support Vector Machine (SVM)	56
4.9.3	Extreme Gradient Boosting (XGBoost)	58
4.9.4	Fine-tuning BERT	59
4.9.5	Hyperparameter tuning	62
4.9.6	Validasi model	64
4.10	Evaluasi Model	64
BAB V IMPLEMENTASI		68
5.1	Pengumpulan Data	68
5.2	Label Mapping	71
5.3	Split Dataset	74
5.4	Pra Pemrosesan Data	75
5.4.1	Case folding	75
5.4.2	Special character removal	75
5.4.3	Normalisasi teks	76
5.4.4	Data augmentation	77
5.4.5	Tokenisasi	79
5.5	Modelling	81
5.5.1	Logistic Regression	81
5.5.2	Support Vector Machine (SVM)	82
5.5.3	Extreme Gradient Boosting (XGBoost)	83
5.5.4	Pembuatan dataset pytorch	83
5.5.5	Inisialisasi konfigurasi model BERT	85
5.5.6	Hyperparameter tuning model BERT	87
5.6	Metrik Evaluasi	88
BAB VI HASIL DAN PEMBAHASAN		91
6.1	Pengumpulan Data	91

6.2 Split Dataset	95
6.3 Hasil Pra-pemrosesan Data	96
6.4 Hasil Data Augmentasi	98
6.5 Hasil Modelling.....	103
6.6 Pengujian Model	113
BAB VII PENUTUP	118
7.1 Kesimpulan	118
7.2 Saran.....	118
DAFTAR PUSTAKA	120

DAFTAR GAMBAR

<u>Gambar 3.1 Arsitektur transformer (Vawsani, 2017)</u>	22
<u>Gambar 3.2 FFN pada BERT Base</u>	25
<u>Gambar 3.3 Arsitektur BERT (Devlin et al., 2019)</u>	25
<u>Gambar 3.4 Representasi input BERT (Devlin, et al., 2019)</u>	26
<u>Gambar 3.5 Multihead attention pada transformer (Vaswani, et al., 2017)</u>	28
<u>Gambar 3.6 Arsitektur klasifikasi multi-label IndoBERT</u>	32
<u>Gambar 4.1 Diagram penelitian</u>	40
<u>Gambar 4.2 Rancangan prosedur pemodelan</u>	43
<u>Gambar 4.3 Pra-pemrosesan data</u>	48
<u>Gambar 4.4 Arsitektur IndoBERT</u>	52
<u>Gambar 4.5 Arsitektur fine-tuning BERT</u>	61
<u>Gambar 5.1 Kode program loading dataset</u>	69
<u>Gambar 5.2 Contoh hasil tes kepribadian</u>	69
<u>Gambar 5.3 Sebaran nilai data twitter</u>	70
<u>Gambar 5.4 Kode program label mapping</u>	73
<u>Gambar 5.5 Kode program split dataset</u>	74
<u>Gambar 5.6 Kode program case folding</u>	75
<u>Gambar 5.7 Kode program special character removal</u>	76
<u>Gambar 5.8 Contoh potongan Indonesian Colloquial Lexicon Dataset</u>	76
<u>Gambar 5.9 Kode program normalisasi</u>	77
<u>Gambar 5.10 Kode program augmentasi data semi-kontekstual</u>	78
<u>Gambar 5.11 Kode program augmentasi data kontekstual</u>	79
<u>Gambar 5.12 Kode program tokenisasi TF-IDF</u>	80
<u>Gambar 5.13 Kode program tokenisasi BERT</u>	80
<u>Gambar 5.14 Kode program logistic regression</u>	82
<u>Gambar 5.15 Kode program SVM</u>	82
<u>Gambar 5.16 Kode program XGBoost</u>	83
<u>Gambar 5.17 Kode program pembuatan dataset pytorch</u>	85
<u>Gambar 5.18 Kode program konfigurasi model BERT</u>	86
<u>Gambar 5.19 Contoh kode konfigurasi hyperparameter tuning model BERT</u>	88
<u>Gambar 5.20 Kode program metrik evaluasi</u>	90
<u>Gambar 6.1 Boxplot sebaran data uji per label</u>	92
<u>Gambar 6.2 Histogram sebaran data uji per label</u>	93
<u>Gambar 6.3 Matriks co-occurrence tiap label</u>	94
<u>Gambar 6.4 Wordcloud non-kontekstual per label</u>	100
<u>Gambar 6.5 Wordcloud semi-kontekstual per label</u>	101
<u>Gambar 6.6 Wordcloud kontekstual per label</u>	103
<u>Gambar 6.7 Feature importance model SVM</u>	108
<u>Gambar 6.8 Feature importance model XGBoost</u>	111

DAFTAR TABEL

Tabel 2.1 Tinjauan pustaka	12
Tabel 3.1 Perbedaan <i>hyperparameter</i> transformer dan BERT	23
Tabel 4.1 Contoh data esai	44
Tabel 4.2 Distribusi data	45
Tabel 4.3 Skenario pengujian performa <i>data augmentation</i>	47
Tabel 4.4 Contoh <i>case folding</i>	49
Tabel 4.5 Contoh <i>special character removal</i>	49
Tabel 4.6 Contoh tokenisasi	49
Tabel 4.7 Contoh <i>stopword removal</i>	50
Tabel 4.8 Contoh nilai <i>hyperparameter</i> yang digunakan	63
Tabel 4.9 Ilustrasi skenario pengujian <i>hyperparameter tuning</i>	65
Tabel 4.10 Skenario pengujian performa model	66
Tabel 5.1 Distribusi kelas data latih	71
Tabel 5.2 Statistika deskriptif panjang teks dataset	71
Tabel 5.3 Penyelarasan label mapping	72
Tabel 5.4 Distribusi jumlah label per <i>instance</i>	73
Tabel 6.1 Sebaran label <i>trait</i> kepribadian	92
Tabel 6.2 Sebaran <i>split dataset</i> label <i>trait</i> kepribadian	95
Tabel 6.3 Potongan hasil dari translasi	96
Tabel 6.4 Potongan hasil dari pra-pemrosesan	96
Tabel 6.5 Potongan hasil dari tokenisasi	97
Tabel 6.6 Hasil pengujian <i>logistic regression</i>	104
Tabel 6.7 Hasil pengujian SVM	105
Tabel 6.8 Hasil pengujian XGBoost	109
Tabel 6.9 Hasil pengujian <i>hyperparameter tuning</i>	111
Tabel 6.10 Perbandingan hasil pengujian model dengan data uji	112