DAFTAR ISI

ABSTRAK i

ABSTRACT ii

KATA PENGANTAR iii

DAFTAR ISI iv

DAFTAR TABEL viii

DAFTAR GAMBAR xii

DAFTAR PERSAMAAN xiv

DAFTAR KODE xv

BAB I PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Rumusan Masalah 3

1.3 Batasan Masalah 3

1.4 Tujuan 4

1.5 Manfaat 4

1.6 Kerangka Pemikiran 4

1.7 Sistematis Penulisan 5

BAB II TINJAUAN PUSTAKA 7

2.1 Hasil Penelitian Terdahulu 7

2.2 Dasar Teori 15

2.2.1 Situs Phising Web3 15

2.2.2 URL dan Domain 19

2.2.3 Konten HTML 20

2.2.4 *Pre Trained* Domsurl BERT 22

2.2.5 *Hugging Face Transformers* 24

2.2.6 *Extreme Gradient boosting* (XGBoost) 29

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
B A N D U N G

2.2.7 Dataset *Split* 32
2.2.8 *Confusion matrix* 33
2.2.9 Metrik Evaluasi 35
2.2.10 *Feature importance* 38
2.2.11 *Web Scraping* 39
2.2.12 *Browser Extension* 40
2.2.13 CRISP-DM (*Cross-Industry Standard Process for Data Mining*) 40
BAB III METODOLOGI 41
3.1 *Bussiness Understanding* 42
3.1.1 Tujuan Bisnis 42
3.1.2 Penilaian Situasi 43
3.1.3 Tujuan Data Mining 43
3.1.4 Rencana Proyek 44
3.2 *Data Understanding* 45
3.2.1 Sumber Dataset Domain 46
3.2.2 Sumber Dataset HTML 48
3.2.3 Sumber Dataset Klasifikasi Akhir 49
3.2.4 Ringkasan Informasi Sumber Dataset 51
3.2.5 Kualitas dan Karakteristik Data 51
3.3 *Data Preparation* 52
3.3.1 *Data Preparation* Dataset Domain 52
3.3.2 Data Preparation Dataset Fitur HTML 56
3.3.3 Data Preparation Dataset Klasifikasi Akhir 67
3.4 Modelling 68
3.4.1 Pemodelan Analisis Domain 69
3.4.2 Pemodelan Analisis Konten HTML 72
3.4.3 Pemodelan Klasifikasi Akhir 76
3.5 *Evaluation* 81
3.6 *Deployment* 81
3.6.1 Rancangan *Deployment* 81
3.6.2 Rancangan Alur Sistem Deteksi 83

UIN SUNAN GUNUNG DJATI BANDUNG

BAB IV HASIL DAN PEMBAHASAN 85
4.1 Hasil *Business Understanding* 85
4.2 Hasil *Data Understanding* 85
4.2.1 Hasil Data Understanding Dataset Domain 86
4.2.2 Hasil Data Understanding Dataset HTML 87
4.2.3 Hasil Data Understanding Dataset klasifikasi akhir 88
4.3 Hasil *Data Preparation* 89
4.2.1 Hasil Data Preparation Dataset Domain 89
4.2.2 Hasil *Data Preparation* Dataset HTML 91
4.2.3 Hasil *Data Preparation* Dataset Klasifikasi Akhir 95
4.2.4 Ringkasan Dataset dan data *split* 96
4.4 Hasil Modelling 97
4.3.1 Hasil Pemodelan Analisis Domain 98
4.3.2 Hasil Pemodelan Analisis fitur HTML 103
4.3.3 Hasil Pemodelan Model Klasifikasi Akhir 105
4.5 Hasil *Evaluation* 110
4.4.1 Hasil *Evaluation* Model Domurls-Bert 110
4.4.2 Hasil *Evaluation* Model HTML 111
4.4.3 Hasil *Evaluation* Model XGboost Klasifikasi Akhir 114
4.6 Hasil *Deployment* 119
4.1.1 Hasil Rancangan *Deployment* 119
4.1.2 Skenario Instalasi Ekstensi oleh Pengguna 128
4.5.1 Antarmuka Ekstensi *Browser* 129
4.1.3 Hasil Pengujian *Black Box* 133
4.1.4 Hasil Pengujian Non-Fungsional 134
4.1.5 Arsitektur Sistem *Deployment* 135
4.1.6 Analisis Hasil Prediksi pada Situs Phising 137
BAB V SIMPULAN DAN SARAN 140
5.1 Simpulan 140
5.1.1 Terhadap Rumusan Masalah 1 140
5.1.2 Simpulan Terhadap Rumusan Masalah 2 140

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

5.2 Saran 141

DAFTAR PUSTAKA 142

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

DAFTAR TABEL

Tabel 2. 1 State Of The Art 7
Tabel 2. 2 Rumus Metrik Evaluasi Berdasarkan *Confusion matrix* 34
Tabel 3. 1 Metadata Dataset PhreshPhish 48
Tabel 3. 2 Status Keaktifan Domain *Scam*Sniffer 49
Tabel 3. 3 Status Keaktifan Domain DefiLlama 49
Tabel 3. 4 Status Keaktifan Domain CryptoRank 50
Tabel 3. 5 Status Keaktifan Domain DataForSEO 50
Tabel 3. 6 Ringkasan Sumber Dataset Klasifikasi Akhir (Domain Aktif) 50
Tabel 3. 7 Ringkasan Sumber Dataset yang Digunakan dalam Penelitian 51
Tabel 3. 8 Teknik Augmentasi dan Contoh Transformasi 53
Tabel 3. 9 Kategori Kata Kunci yang Digunakan dalam Augmentasi 54
Tabel 3. 10 Contoh Hasil Ekstraksi *Root domain* 55
Tabel 3. 11 Pembagian Dataset A (Domain) 56
Tabel 3. 12 Kondisi yang Dikategorikan sebagai Empty Link 57
Tabel 3. 13 Kriteria Penentuan Fitur title_contains_domain 59
Tabel 3. 14 Contoh Pola Base64 Data URI yang Dihitung pada Fitur num_base64_data_uri 60
Tabel 3. 15 Sumber teks yang diperiksa pada ekstraksi copyright_contains_domain 61
Tabel 3. 16 Kriteria pencocokan domain pada copyright_contains_domain 62
Tabel 3. 17 Kata Kunci yang Digunakan untuk Mendeteksi Tautan Terms dan Privacy 64
Tabel 3. 18 Contoh Perhitungan Fitur css_important_count 64
Tabel 3. 19 Mekanisme Deteksi Fitur *wallet*_connect_popup 64
Tabel 3. 20 Pembagian Dataset B (fitur HTML) 67
Tabel 3. 21 Pembagian Dataset C (klasifikasi Akhir) 68
Tabel 3. 22 *Hyperparameter* Optimasi Pemodelan Domain 71
Tabel 3. 23 *Hyperparameter* Regularisasi dan Pelatihan Pemodelan Domain 71
Tabel 3. 24 Fitur yang Digunakan dalam Pelatihan Model HTML XGBoost 73

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Tabel 3. 25 *Hyperparameter* Pemodelan HTML (n2) 75
Tabel 3. 26 Konfigurasi Pelatihan Model XGBoost HTML.....75
Tabel 3. 27 Fitur Probabilitas Inferensi dan HTML 79
Tabel 3. 28 *Hyperparameter* Model Klasifikasi Akhir XGBoost (n3) 80
Tabel 3. 29 Konfigurasi Pelatihan pada Model Klasifikasi Akhir XGBoost (n3) .80
Tabel 4. 1 Head DataForSEO (1000 baris) 86
Tabel 4. 2 Head CryptoRank (3.062 baris) 86
Tabel 4. 3 Head DeFiLlama (5053 baris).....87
Tabel 4. 4 Head *Scam*Sniffer (11.000 baris) 87
Tabel 4. 5 Head Dataset HTML Sebelum Ekstraksi Fitur (199989 baris).....87
Tabel 4. 6 Head Dataset Klasifikasi Akhir Kelas *Legitimate* Sebelum Ekstraksi HTML (6.158 baris) 88
Tabel 4. 7 Head Dataset Klasifikasi Akhir Kelas *Phishing* Sebelum Ekstrak HTML (5.200 baris) 88
Tabel 4. 8 Komposisi Dataset Domain Setelah Penggabungan 89
Tabel 4. 9 Ringkasan Hasil Pembersihan Data 89
Tabel 4. 10 Distribusi Label Dataset Sebelum dan Sesudah Pembersihan Data ...90
Tabel 4. 11 Ringkasan Hasil Augmentasi Dataset *Phishing*.....90
Tabel 4. 12 Distribusi Label Dataset Sebelum dan Sesudah Augmentasi 90
Tabel 4. 13 Ringkasan Hasil Ekstraksi *Root domain* pada Dataset Legit.....90
Tabel 4. 14 Distribusi Label Dataset Sebelum dan Sesudah Ekstraksi *Root domain* 91
Tabel 4. 15 Distribusi Kelas Dataset Domain Setelah Pembersihan 91
Tabel 4. 16 Head Dataset Domain Setelah Normalisasi 91
Tabel 4. 17 Daftar Fitur Hasil Ekstraksi dari Konten HTML 92
Tabel 4. 18 Ringkasan Tahapan Pembersihan Dataset HTML 92
Tabel 4. 19 Distribusi Label Dataset HTML Sebelum dan Sesudah Pembersihan93
Tabel 4. 20 Hasil Perhitungan Variance Fitur.....93
Tabel 4. 21 Distribusi Kelas pada Dataset HTML 94
Tabel 4. 22 Head Dataset HTML Setelah Ekstraksi Fitur 94
Tabel 4. 23 Head Dataset Klasifikasi Akhir Setelah Ekstraksi Fitur 95

Tabel 4. 24 Ringkasan Tahapan Pembersihan Dataset Klasifikasi Akhir.....95
Tabel 4. 25 Distribusi Label Dataset Klasifikasi Akhir Sebelum dan Sesudah Pembersihan 96
Tabel 4. 26 Distribusi Kelas Akhir Dataset Klasifikasi Akhir.....96
Tabel 4. 27 Distribusi Kelas Dataset A (Domain) 96
Tabel 4. 28 Distribusi Kelas Dataset B (Fitur HTML) 97
Tabel 4. 29 Distribusi Kelas Dataset C (Fitur Gabungan Domain dan HTML) 97
Tabel 4. 30 Statistik Panjang Token Dataset 98
Tabel 4. 31 Perbandingan Metrik Evaluasi Berdasarkan Rasio *Split* Dataset.....98
Tabel 4. 32 Perbandingan Metrik Evaluasi Berdasarkan Learning Rate 99
Tabel 4. 33 Perbandingan Metrik Evaluasi Berdasarkan Batch Size.....99
Tabel 4. 34 Perbandingan Metrik Evaluasi Berdasarkan Warmup Ratio 99
Tabel 4. 35 Perbandingan Metrik Evaluasi Berdasarkan LR Scheduler.....100
Tabel 4. 36 Perbandingan Metrik Evaluasi Berdasarkan Weight Decay 100
Tabel 4. 37 Perbandingan Metrik Evaluasi Berdasarkan *Epoch* dan Early Stopping 100
Tabel 4. 38 *Hyperparameter* Final yang dipilih 100
Tabel 4. 39 Hasil Evaluasi pada Setiap *Epoch* Fine-tuning Domurls-Bert.....102
Tabel 4. 40 Artifact Model DomURLs-BERT yang Diperlukan untuk *Deployment* 102
Tabel 4. 41 Hasil Eksperimen *Split* Dataset XGBoost HTML 103
Tabel 4. 42 Hasil *Hyperparameter* Terbaik XGBoost HTML.....104
Tabel 4. 43 Hasil Penentuan n_estimators Menggunakan Early Stopping 104
Tabel 4. 44 Artifacts Model XGBoost untuk Klasifikasi Konten HTML.....105
Tabel 4. 45 Head Nilai Probabilitas *prob_domain* dan *prob_html* 105
Tabel 4. 46 *head* Fitur Hasil Kombinasi Probabilitas Model Dasar 106
Tabel 4. 47 Hasil Analisis Varians Fitur Gabungan pada Data Latih.....107
Tabel 4. 48 Perbandingan Metrik Evaluasi Model Klasifikasi Akhir Berdasarkan Rasio *Split* Dataset 107
Tabel 4. 49 Konfigurasi Final Model Klasifikasi Akhir Terpilih 108
Tabel 4. 50 Hasil Penentuan n_estimators Model Klasifikasi Akhir Menggunakan

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

Early Stopping.....109
Tabel 4. 51 Artifact Model Klasifikasi Akhir yang Diperlukan untuk Deployment 109
Tabel 4. 52 Hasil Evaluasi Model DomURLs-BERT pada *Testing* Dataset A....111
Tabel 4. 53 Hasil Evaluasi Model XGBoost HTML pada *Testing Dataset B* 112
Tabel 4. 54 Hasil Evaluasi Model Klasifikasi Akhir pada *Testing Dataset C*.....114
Tabel 4. 55 Perbandingan Performa Seluruh Model (Test Set) 116
Tabel 4. 56 Contoh Sampel dengan Nilai Disagreement Tinggi 118
Tabel 4. 57 Spesifikasi Endpoint API Inferensi.....125
Tabel 4. 58 Daftar Halaman Antarmuka Ekstensi *Browser* dan Fungsinya 131
Tabel 4. 59 Hasil Pengujian *Black Box* Ekstensi Browser.....133
Tabel 4. 60 Hasil Pengujian Performa Waktu Respons Sistem 134

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

DAFTAR GAMBAR

Gambar 1. 1 Kerangka Pemikiran 5
Gambar 2. 1 Contoh Tampilan Website *Phishing* pada Skema *Airdrop* 17
Gambar 2. 2 Contoh Tampilan Website *Phishing* yang Meniru Platform DeFi 17
Gambar 2. 3 Struktur URL 19
Gambar 2. 4 Arsitektur Pre-training DomURLs-BERT 24
Gambar 2. 5 Ilustrasi Proses Three-Way Dataset Split 32
Gambar 2. 6 Ilustrasi Proses 5-Fold Cross-Validation 33
Gambar 2. 7 Struktur *Confusion matrix* pada Klasifikasi Biner 34
Gambar 2. 8 Kurva ROC AUC 36
Gambar 2. 9 Tahapan Metodologi CRISP-DM 40
Gambar 3. 1 Tahapan *Business Understanding*, Data Understanding, dan *Data Preparation* pada Metodologi CRISP-DM 41
Gambar 3. 2 Tahapan *Modeling*, *Evaluation*, dan *Deployment* pada Metodologi CRISP-DM 42
Gambar 3. 3 Proses Ekstraksi *Root domain* dari URL 57
Gambar 3. 4 Diagram Alur Proses Fine-Tuning Model DomURLs-BERT 69
Gambar 3. 5 Diagram Alur Modelling XGBoost 74
Gambar 3. 6 Diagram Alur Pemodelan XGBoost Klasifikasi Akhir 77
Gambar 3. 7 Diagram Sequence Alur Sistem Deteksi 84
Gambar 4. 1 Heatmap Korelasi Antar Fitur pada Dataset HTML 94
Gambar 4. 2 *Learning Curve* Proses Pelatihan DomURLs-BERT 101
Gambar 4. 3 Matriks Korelasi Antar Fitur Model Klasifikasi Akhir 106
Gambar 4. 4 *Confusion matrix* DomURLs-BERT pada Data Uji 110
Gambar 4. 5 *Confusion matrix* XGBoost HTML pada Data Uji 112
Gambar 4. 6 Peringkat *Feature importance* pada Model XGBoost HTML 113
Gambar 4. 7 Peringkat Feature importance Sepuluh Fitur Teratas pada Model XGBoost Klasifikasi Akhir 115
Gambar 4. 8 Perbandingan *Confusion matrix* Empat Model Klasifikasi pada Dataset C (Testing) 117

Gambar 4. 9 Hasil Ekstraksi Fitur dalam Format JSON 120
Gambar 4. 10 Hasil Prediksi Model Klasifikasi Akhir 124
Gambar 4. 11 Hasil Response API pada Postman 127
Gambar 4. 12 Ekstensi PhishPredict di Chrome Web Store 129
Gambar 4. 13 Tampilan Hasil Deteksi *Phishing*.....129
Gambar 4. 14 Gambar Contoh Prediksi Situs *Phishing* Pada Extension 130
Gambar 4. 15 Contoh Deteksi Website Secara Real time (kasus *Phishing*).....131
Gambar 4. 16 Contoh Deteksi Website Secara Real time (kasus Legit).....132
Gambar 4. 17 Arsitektur Sistem Deployment.....135
Gambar 4. 18 Hasil Prediksi Pada Ekstensi.....137

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

DAFTAR PERSAMAAN

Persamaan 2. 1 Perhitungan *Accuracy* 35
Persamaan 2. 2 Perhitungan Error Rate 35
Persamaan 2. 3 Perhitungan *Precision* 36
Persamaan 2. 4 Perhitungan *Recall* 36
Persamaan 2. 5 Perhitungan F1 Score 36
Persamaan 2. 6 Perhitungan *Macro Average Precision* 37
Persamaan 2. 7 Perhitungan *Macro Average Recall* 37
Persamaan 2. 8 Perhitungan *Macro Average F1-Score* 37
Persamaan 2. 9 Perhitungan *Weighted Average Precision* 38
Persamaan 2. 10 Perhitungan *Weighted Average Recall* 38
Persamaan 2. 11 Perhitungan *Weighted Average F1-Score* 38

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG

DAFTAR KODE

Kode 2. 1 Pemuatan AutoTokenizer dari Model Pra-latih.....25
Kode 2. 2 Pemuatan Konfigurasi Model Menggunakan AutoConfig.....25
Kode 2. 3 Pemuatan Model Klasifikasi Menggunakan AutoModelForSequence Classification.....26
Kode 2. 4 Konfigurasi Parameter Pelatihan Menggunakan TrainingArguments ..26
Kode 2. 5 Konfigurasi EarlyStoppingCallback.....29
Kode 2. 6 Inisialisasi Trainer 29
Kode 2. 7 Inisialisasi Model XGBoost Menggunakan XGBClassifier.....30
Kode 4. 1 Kode Program Memuat Artifacts Hasil Modelling 121
Kode 4. 2 Kode Program Memuat Data Masukan 121
Kode 4. 3 Kode Program Praproses dan Prediksi Model XGBoost HTML 122
Kode 4. 4 Kode Program Prediksi Model DomURLs-BERT 123
Kode 4. 5 Kode Program Pembentukan Fitur Gabungan.....123
Kode 4. 6 Kode Program Prediksi Hasil Klasifikasi Akhir 124
Kode 4. 7 Konfigurasi manifest.json Ekstensi Browser 127
Kode 4. 8 Hasil Pencarian Ekstensi PhishPredict 128

UNIVERSITAS ISLAM NEGERI
SUNAN GUNUNG DJATI
BANDUNG