



SKRIPSI

**KLASIFIKASI SERTIFIKAT BERDASARKAN
MATA KULIAH DALAM PROGRAM REKOGNISI
PEMBELAJARAN LAMPAU
BERBASIS NLP MENGGUNAKAN BERT**

DIMAS SAPUTRA
NPM 21081010151

DOSEN PEMBIMBING

Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU
Eva Yulia Puspaningrum, S.Kom., M.Kom

**KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2025**

LEMBAR PENGESAHAN

KLASIFIKASI SERTIFIKAT BERDASARKAN MATA KULIAH DALAM PROGRAM REKOGNISI PEMBELAJARAN LAMPAU BERBASIS NLP MENGGUNAKAN BERT

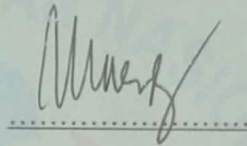
Oleh :
DIMAS SAPUTRA
NPM. 21081010151

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada tanggal 27 Mei 2025

Menyetujui

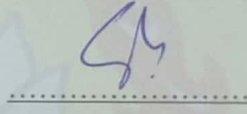
Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT.
IPU

NIP. 19700619 202121 1 009



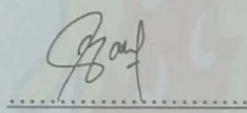
(Pembimbing I)

Eva Yulia Puspaningrum, S.Kom., M.Kom
NIP. 19890705 202121 2 002



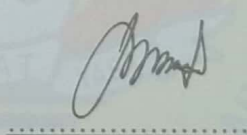
(Pembimbing II)

Made Hanindia Prami Swari, S.Kom, M.Cs
NIP. 19890205 201803 2 001



(Ketua Penguji)

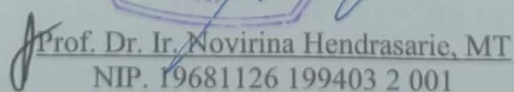
Achmad Junaidi, S.Kom., M.Kom
NPT. 3 7811 04 0199 1



(Anggota Penguji)

Mengetahui,
Dekan Fakultas Ilmu Komputer




Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

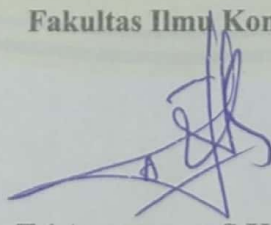
LEMBAR PERSETUJUAN

**KLASIFIKASI SERTIFIKAT BERDASARKAN MATA KULIAH DALAM
PROGRAM REKOGNISI PEMBELAJARAN LAMPAU
BERBASIS NLP MENGGUNAKAN BERT**

Oleh :
DIMAS SAPUTRA
NPM. 21081010151

Menyetujui,

**Kooridnator Program Studi Informatika
Fakultas Ilmu Komputer**



Fetty Tri Anggraeny, S.Kom., M.Kom.

NIP. 19820211 202121 2 005

SURAT PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini:

Nama : Dimas Saputra
NPM : 21081010151
Program : Sarjana (S1)
Program Studi : Informatika
Fakultas : Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila dikemudian hari ditemukan indikasi plagiat pada Skripsi ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.



Surabaya, 27 Mei 2025
Yang Membuat Pernyataan,

Dimas



Dimas Saputra
NPM. 21081010151

ABSTRAK

Nama Mahasiswa / NPM : Dimas Saputra / 21081010151
Judul Skripsi : Klasifikasi Sertifikat Berdasarkan Mata Kuliah
Dalam Program Rekognisi Pembelajaran Lampau
Berdasarkan NLP Menggunakan BERT
Dosen Pembimbing : 1. Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU
2. Eva Yulia Puspaningrum, S.Kom., M.Kom

Perkembangan teknologi informasi yang pesat mendorong kebutuhan akan sistem otomatisasi dalam pengakuan pembelajaran non-formal, salah satunya melalui Rekognisi Pembelajaran Lampau (RPL). Sertifikat kompetensi yang diperoleh dari pelatihan mandiri sering kali tidak terintegrasi secara sistematis ke dalam program akademik. Penelitian ini bertujuan untuk mengembangkan model klasifikasi otomatis yang dapat mengelompokkan sertifikat ke dalam mata kuliah terkait dalam lingkup program studi Informatika. Tujuh kategori mata kuliah ditentukan sebagai kelas target, meliputi: Machine Learning, Pemrograman Web, Desain Antarmuka, Jaringan Komputer, Aplikasi Game, Pemrograman Mobile, dan Manajemen Proyek. Metodologi yang digunakan meliputi ekstraksi data sertifikat berbentuk PDF melalui proses OCR menggunakan PyTesseract, diikuti dengan tahap praproses teks dan data *Augmentation* untuk memperbaiki distribusi kelas. Model yang diimplementasikan adalah BERT (*Bidirectional Encoder Representations from Transformers*) dengan dua konfigurasi, yakni bert-base-uncased dan bert-base-multilingual-uncased. Lima skenario augmentasi diuji, yaitu tanpa augmentasi, *Character Insertion*, *Character Deletion*, *Back Translation*, dan *Synonym Replacement*. Hasil evaluasi menunjukkan bahwa konfigurasi bert-base-uncased dengan teknik augmentasi *Synonym Replacement* menghasilkan performa terbaik, dengan akurasi validasi mencapai 95,54% dan *F1-score* sebesar 0,97. Temuan ini menegaskan efektivitas BERT dalam klasifikasi teks berbahasa Indonesia-Inggris dan manfaat augmentasi semantik dalam meningkatkan generalisasi model. Sebagai implementasi praktis, penelitian ini juga mengembangkan prototipe layanan berbasis Streamlit yang memungkinkan klasifikasi otomatis sertifikat melalui antarmuka sederhana. Model dan sistem yang dikembangkan diharapkan dapat mendukung integrasi kompetensi non-formal ke dalam kurikulum akademik secara lebih efisien dan akurat.

Kata kunci : Klasifikasi Teks, BERT, Rekognisi Pembelajaran Lampau, Augmentasi Data, Pemrosesan Bahasa Alami

ABSTRACT

Student Name / NPM : Dimas Saputra / 21081010151
Thesis Title : Classification of Certificates Based on Courses in
Prior Learning Recognition Program Using NLP
with BERT
Advisor : 1. Dr. Ir. I Gede Susrama Mas Diyasa, ST. MT. IPU
2. Eva Yulia Puspaningrum, S.Kom., M.Kom

ABSTRACT

The rapid advancement of information technology has increased the demand for automated systems to recognize non-formal learning, particularly through the Recognition of Prior Learning (RPL) program. Competency certificates obtained from independent training programs are often not systematically integrated into academic curricula. This study aims to develop an automated classification model capable of categorizing certificates into relevant course subjects within the Informatics study program. Seven target course categories were defined, including Machine Learning, Web Programming, Interface Design, Computer Networks, Game Applications, Mobile Programming, and Project Management. The methodology includes extracting text from certificate documents in PDF format through Optical Character Recognition (OCR) using PyTesseract, followed by text preprocessing and data Augmentation to improve class distribution. The implemented model is BERT (Bidirectional Encoder Representations from Transformers), evaluated in two configurations: bert-base-uncased and bert-base-multilingual-uncased. Five data scenarios were tested: no Augmentation, Character Insertion, Character Deletion, Back Translation, and Synonym Replacement. Evaluation results indicate that the bert-base-uncased configuration with Synonym Replacement Augmentation yielded the best performance, achieving a validation accuracy of 95.54% and an *F1-score* of 0.97. These findings confirm the effectiveness of BERT for text classification in both Indonesian and English, and highlight the benefit of semantic-based Augmentation techniques in improving model generalization. As a practical implementation, this research also developed a prototype service using the Streamlit framework, enabling automated certificate classification through a user-friendly interface. The model and system developed are expected to support the efficient and accurate integration of non-formal competencies into academic programs.

Keywords: Text Classification, BERT, Recognition of Prior Learning, Data Augmentation, Natural Language Processing

KATA PENGANTAR

Puji syukur kehadiran Allah SWT atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga skripsi dengan judul **“Klasifikasi Sertifikat Berdasarkan Mata Kuliah Dalam Program Rekognisi Pembelajaran Lampau Berbasis NLP Menggunakan BERT”** dapat terselesaikan dengan baik.

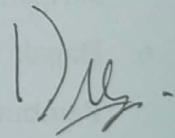
Peneliti menyadari tidak akan dapat menyelesaikan skripsi ini dengan baik tanpa rida, bimbingan, saran, motivasi, dan bantuan dari berbagai pihak. Pada kesempatan ini, penulis mengucapkan terimakasih yang sebesar-besarnya kepada:

1. Kedua orang tua serta kakak, selaku keluarga yang selalu memberikan doa, dukungan moril maupun materiil, serta dukungan yang tiada henti.
2. Bapak Prof. Dr. Ir. Akhmad Fauzi, M.MT. selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
3. Ibu Dr. Ir. Novirina Hendrasarie, MT. selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
4. Ibu Fetty Try Anggraeny, S.Kom, M.Kom. selaku Koordinator Program Studi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
5. Bapak Andreas Nugroho Sihananto, S.Kom., M.Kom. selaku Koordinator Skripsi, yang telah membantu dalam proses administrasi dan pelaksanaan skripsi.
6. Bapak Dr. Ir. I Gede Susrama Mas Diyasa, ST., MT., IPU. selaku Dosen Pembimbing pertama, yang telah sabar memberikan bimbingan, motivasi, dan waktunya dalam proses bimbingan penyusunan skripsi
7. Ibu Eva Yulia Puspaningrum, S.Kom., M.Kom. selaku Dosen Pembimbing kedua, yang telah memberikan arahan dan saran dalam proses penyusunan skripsi.
8. Ibu Made Hanindia Prami Swari, S.Kom., M.Cs., selaku Ketua Penguji dan Dosen Wali, yang telah memberikan bimbingan, nasehat, kritik dan saran yang sangat berharga selama masa perkuliahan dan proses penyusunan skripsi ini.

9. Bapak Achmad Junaidi, S.Kom., M.Kom., selaku Anggota Penguji, yang telah memberikan kritik dan masukan yang sangat bermanfaat dalam proses penyusunan skripsi ini.
10. Seluruh staf Tata Usaha dan civitas akademika Universitas Pembangunan Nasional "Veteran" Jawa Timur atas bantuan administrasi dan pelayanan selama penulis menempuh pendidikan.
11. Seluruh keluarga besar dan kerabat yang senantiasa memberikan doa dan dukungan.
12. Teman-teman seperjuangan yang selalu memberikan semangat, motivasi, dan kebersamaan selama proses penyusunan skripsi.
13. Semua pihak yang tidak dapat disebutkan satu per satu yang telah membantu secara langsung maupun tidak langsung dalam penyusunan skripsi ini.

Penulis menyadari bahwa di dalam penyusunan skripsi ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan skripsi ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya, 27 Mei 2025



Dimas Saputra

DAFTAR ISI

LEMBAR PENGESAHAN	i
LEMBAR PERSETUJUAN	iii
SURAT PERNYATAAN BEBAS PLAGIAT	v
ABSTRAK	vii
ABSTRACT	ix
KATA PENGANTAR.....	xi
DAFTAR ISI.....	xiii
DAFTAR TABEL	xvii
DAFTAR GAMBAR	xix
BAB 1 PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	2
1.3 Tujuan Penelitian	2
1.4 Manfaat Penelitian	3
1.5 Batasan Masalah	3
BAB 2 TINJAUAN PUSTAKA.....	5
2.1 Penelitian Terdahulu	5
2.2 Sertifikat.....	6
2.3 Website Scraping.....	6
2.4 <i>Optical Character Recognition (OCR)</i>	7
2.5 Praproses	7
2.5.1 <i>Case folding</i>	7
2.5.2 Filter Stopword.....	8

2.5.3	<i>Lemmatization</i>	8
2.6	<i>Data Text Augmentation</i>	8
2.6.1	<i>Character Insertion</i>	9
2.6.2	<i>Character Deletion</i>	9
2.6.3	<i>Back Translation</i>	9
2.6.4	<i>Synonym Replacement</i>	9
2.7	Machine Learning	10
2.7.1	Natural Language Processing	10
2.7.2	Artificial Neural Network	10
2.7.3	Softmax	11
2.7.4	Transfer Learning	12
2.7.5	Transformer	12
2.7.6	Optimizer AdamW	13
2.7.7	Loss Function Categorical Cross Entropy	13
2.7.8	Confusion Matrix	13
2.8	BERT (<i>Bidirectional Encoder Representations from Transformers</i>)	14
2.8.1	Tokenisasi dan Embedding	15
2.8.2	Arsitektur BERT	15
2.9	Streamlit	17
BAB 3 DESAIN DAN IMPLEMENTASI SISTEM		19
3.1	Tahapan Penelitian	19
3.2	Pengumpulan Data	20
3.2.1	Web Scraping	23
3.2.2	Kuesioner	24
3.3	Pemformatan Data	26

3.3.1	Konversi PDF ke Gambar	27
3.3.2	<i>Optical Character Recognition</i>	28
3.4	Praproses Data	32
3.4.1	<i>Data Cleaning</i>	33
3.4.2	<i>Case folding</i>	34
3.4.3	Filter Stopword.....	35
3.4.4	<i>Lemmatization / Stemming</i>	37
3.4.5	Tokenisasi.....	38
3.5	Augmentasi Data.....	39
3.5.1	<i>Character Insertion</i>	40
3.5.2	<i>Character Deletion</i>	41
3.5.3	<i>Back Translation</i>	43
3.5.4	Synonym Replacemet.....	44
3.6	Pembagian Data	47
3.7	Visualisasi Data.....	48
3.8	Perancangan dan Pelatihan Model BERT	53
3.9	Evaluasi Hasil	58
3.10	Pengembangan Layanan Sederhana.....	59
BAB 4	PENGUJIAN DAN ANALISA	61
4.1	Skenario Pelatihan dan Matrik Evaluasi	61
4.1	Analisis Akurasi dan Validasi Akurasi	62
4.1.1	Akurasi	62
4.1.1	Validasi Akurasi.....	64
4.2	Waktu Eksekusi Training	67
4.3	Confusion Matrix	68

4.4	Classification Report.....	71
4.5	Hasil Prediksi dalam Layanan Sederhana.....	72
4.6	Komparasi dengan Model Lainnya.....	73
BAB 5 PENUTUP		75
5.1	Kesimpulan	75
5.2	Saran	76
DAFTAR PUSTAKA.....		79
LAMPIRAN.....		85

DAFTAR TABEL

Tabel 3. 1 Kata kunci penelusuran untuk web scraping.....	21
Tabel 3. 2 Kode Konversi PDF menjadi Gambar	28
Tabel 3. 3 Kode OCR.....	29
Tabel 3.4 Parameter PyTesseract	29
Tabel 3. 5 Proses <i>Data Cleaning</i>	34
Tabel 3. 6 Proses <i>Case folding</i>	35
Tabel 3. 7 Daftar stopwords pada tiap bahasa	36
Tabel 3. 8 Proses Filter Stopword	36
Tabel 3. 9 Daftar Kata Lemma dan Stem.....	37
Tabel 3. 10 Proses <i>Lemmatization/Stemming</i>	37
Tabel 3. 11 BERT Vocabulary Token	38
Tabel 3. 12 Proses Tokenisasi	39
Tabel 3. 13 Kode Augmentasi <i>Character Insertion</i>	40
Tabel 3. 14 Proses <i>Character Insertion</i>	41
Tabel 3. 15 Kode Augmentasi <i>Character Deletion</i>	42
Tabel 3. 16 Proses <i>Character Deletion</i>	42
Tabel 3. 17 Kode <i>Back Translation</i>	43
Tabel 3. 18 Proses <i>Back Translation</i>	44
Tabel 3. 19 Daftar Kata dan Sinonimnya.....	45
Tabel 3. 20 Kode Augmentasi <i>Synonym Replacement</i>	45
Tabel 3. 21 Proses <i>Synonym Replacement</i>	46
Tabel 3. 22 Rasio Data Latih dan Data Uji	47
Tabel 3. 23 Kode Pelatihan Model BERT.....	56
Tabel 3. 24 Nilai Hyperparameter.....	57
Tabel 4. 1 Skenario Pengujian	61
Tabel 4. 2 Akurasi, Validasi Akurasi, Loss, Validasi Loss Pelatihan Model.....	65
Tabel 4. 3 Classification Report.....	71

DAFTAR GAMBAR

Gambar 2. 1 Ilustrasi Artificial Neural Network	11
Gambar 2. 2 Multi-class Confusion Matrix	14
Gambar 2. 3 Embedding pada BERT	15
Gambar 2. 4 Arsitektur BERT	16
Gambar 2. 5 Pre-training dan Fine-tuning pada BERT	16
Gambar 3. 1 Tahapan Penelitian	19
Gambar 3.2 Tahapan Pengumpulan Data	21
Gambar 3. 3 Google <i>Query</i> Result Dokumen Sertifikat	23
Gambar 3. 4 Kuisioner Penelitian	24
Gambar 3.5 Sampel Data Mentah Sertifikat bentuk PDF	26
Gambar 3. 6 Tahapan Pemformatan Data	27
Gambar 3. 7 Contoh Hasil Penggunaan OCR pada Dokumen	30
Gambar 3. 8 Rancangan Struktur Folder Data Sertifikat	31
Gambar 3. 9 Proses Pemformatan Data	32
Gambar 3.10 Tahapan Praproses	33
Gambar 3. 11 Tahapan Augmentasi	40
Gambar 3. 12 Proporsi Label Data	49
Gambar 3. 13 Proporsi Bahasa pada Data	50
Gambar 3. 14 Distribusi Perbandingan Data asli dan Data Augmentasi	51
Gambar 3. 15 Jumlah Kata pada tiap Dokumen	52
Gambar 3. 16 Kata yang sering muncul	52
Gambar 3. 17 WordCloud	53
Gambar 3. 18 Tahapan Pelatihan Model	54
Gambar 3. 19 Detail Tahapan Pelatihan dalam Model	55
Gambar 3. 20 Tampilan Layanan Sederhana	60
Gambar 4. 1 Grafik Akurasi BERT-base-uncased	63
Gambar 4. 2 Grafik Akurasi BERT-base-multilingual-uncased	63
Gambar 4. 3 Grafik Validasi Akurasi BERT-base-uncased	64
Gambar 4. 4 Grafik Validasi Akurasi BERT-base-multilingual-uncased	65

Gambar 4. 5 Ekskusi Waktu Pelatihan Model BERT-base-uncased	67
Gambar 4. 6 Ekskusi Waktu Pelatihan Model BERT-base-multilingual-uncased	68
Gambar 4. 7 Confusion Matrix Model BERT-base-uncased	69
Gambar 4. 8 Confusion Matrix Model BERT-base-multilingual-uncased	70
Gambar 4. 9 Hasil Prediksi Label Sertifikat dalam Layanan Sederhana	73
Gambar 4. 10 Komparasi Akurasi dan Validasi Akurasi pada berbagai model	74