

**IMPLEMENTASI ALGORITMA LOGISTIC REGRESSION UNTUK
KLASIFIKASI TEKS PADA JUDUL PENELITIAN DOSEN
UNIVERSITAS MULTIMEDIA NUSANTARA**



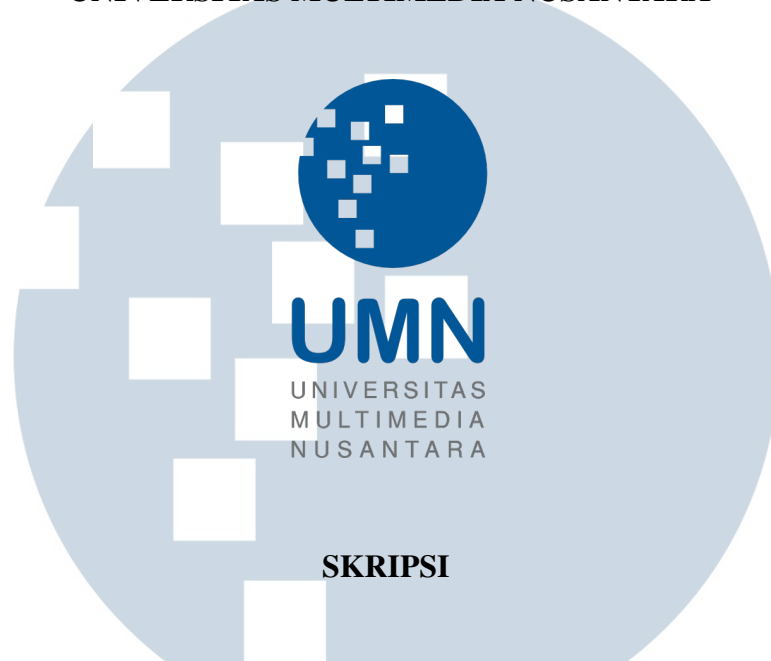
SKRIPSI

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

Felix Ferdianto
00000045398

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2024**

**IMPLEMENTASI ALGORITMA LOGISTIC REGRESSION UNTUK
KLASIFIKASI TEKS PADA JUDUL PENELITIAN DOSEN
UNIVERSITAS MULTIMEDIA NUSANTARA**



SKRIPSI

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

Felix Ferdianto

00000045398

UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2024

HALAMAN PERNYATAAN TIDAK PLAGIAT

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Felix Ferdianto
Nomor Induk Mahasiswa : 00000045398
Program Studi : Informatika

Skripsi dengan judul:

Implementasi Algoritma Logistic Regression Untuk Klasifikasi Teks Pada Judul Penelitian Dosen Universitas Multimedia Nusantara

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas akhir yang telah saya tempuh.

UMMN
UNIVERSITAS

Tangerang, 13 Mei 2024



(Felix Ferdianto)

UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PENGESAHAN

HALAMAN PENGESAHAN

Skripsi dengan judul

**IMPLEMENTASI ALGORITMA LOGISTIC REGRESSION UNTUK
KLASIFIKASI TEKS PADA JUDUL PENELITIAN DOSEN
UNIVERSITAS MULTIMEDIA NUSANTARA**

oleh

Nama : Felix Ferdianto
NIM : 00000045398
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Selasa, 04 Juni 2024

Pukul 08.00 s/s 10.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

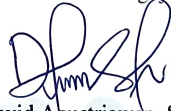
Ketua Sidang



(Alexander Waworuntu, S.Kom., M.T.I.)

NIDN: 0309068503

Penguji

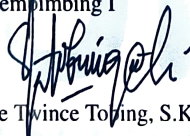


(David Agustriawan, S.Kom., M.Sc.,

Ph.D.)

NIDN: 0525088601

Pembimbing I



(Fenina Adline Twince Toing, S.Kom., M.Kom.)

NIDN: 0406058802

Pembimbing II



(Eunike Endariahna Surbakti, S.Kom., M.T.I.)

NIDN: 0322099401

PJS Ketua Program Studi Informatika,



(Dr. Eng. Niki Prastomo, S.T., M.Sc.)

NIDN: 0419128203

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Felix Ferdianto
NIM : 00000045398
Program Studi : Informatika
Jenjang : S1
Jenis Karya : Skripsi

Menyatakan dengan sesungguhnya bahwa:

- Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya di repositori Knowledge Center, sehingga dapat diakses oleh Civitas Akademika/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial dan saya juga tidak akan mencabut kembali izin yang telah saya berikan dengan alasan apapun.
- Saya tidak bersedia karena dalam proses pengajuan untuk diterbitkan ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*)**.

Tangerang, 13 Mei 2024

Yang menyatakan



Felix Ferdianto

U M M N
U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

** Jika tidak bisa membuktikan LoA jurnal/HKI selama enam bulan ke depan, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN.

Halaman Persembahan / Motto

"If it looks wrong but it works, it isn't wrong"

Mercedes Lackey

"The second you stop caring about a people like you, you'll be successful."

"Is it better to be hated for what you are, than loved for what you are not?."

"The Chainsmokers - Somebody ft. Drew Love (Rory Kramer Vision)"
(Video YouTube)

"Jika ada yang kau pelajari dari masa lalumu
Jaga apinya, hadapi hari, teruslah kau begitu
Jika ada pedih yang panjang mengikat tubuhmu
Percayalah, dunia tak selamanya harus begitu"

Lomba Sihir - Ribuan Memori

UMMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: Implementasi Algoritma Logistic Regression Untuk Klasifikasi Teks Pada Judul Penelitian Dosen Universitas Multimedia Nusantara dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi saya untuk menyelesaikan skripsi ini. Oleh karena itu, saya mengucapkan terima kasih kepada:

1. Bapak Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku PJS Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Ibu Fenina Adline Twince Tobing, S.Kom., M.Kom., sebagai Pembimbing pertama yang telah banyak meluangkan waktu untuk memberikan bimbingan, arahan dan motivasi atas terselesainya Skripsi ini.
5. Ibu Eunike Endariahna Surbakti, S.Kom., M.T.I., sebagai Pembimbing kedua yang telah banyak membantu dan memberikan bimbingan atas terselesainya Skripsi ini.
6. Orang Tua dan teman-teman saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan skripsi ini.

Semoga skripsi ini bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 13 Mei 2024



Felix Ferdianto

**IMPLEMENTASI ALGORITMA LOGISTIC REGRESSION UNTUK
KLASIFIKASI TEKS PADA JUDUL PENELITIAN DOSEN
UNIVERSITAS MULTIMEDIA NUSANTARA**

Felix Ferdianto

ABSTRAK

Lembaga Penelitian dan Pengabdian Masyarakat Universitas Multimedia Nusantara (LPPM UMN) mengelola pengelompokan judul penelitian dosen UMN secara manual, melibatkan *screening* 3000 lebih judul penelitian milik dosen UMN dari tahun 2018-2023 tanpa sistem tambahan. Untuk meningkatkan efisiensi, dibangun sistem klasifikasi teks menggunakan algoritma *logistic regression*. Algoritma ini memodelkan hubungan antara hasil biner dan variabel prediktor, menggunakan penalti L1 dan solver saga. Model ini memanfaatkan *library data preprocessing* TF-IDF (*Term Frequency Inverse Document Frequency*) yang merupakan *library* untuk menghitung frekuensi pentingnya setiap kata yang muncul pada *dataset*. Model ini juga menggunakan dataset yang tanpa melewati proses *data cleaning* dengan menggunakan pembagian *dataset* 80% untuk *training* dan 20% untuk *testing*. Hasil uji coba menunjukkan model ini merupakan model terbaik dengan akurasi 90.01%, presisi 0.86, *recall* 0.82, dan *f1-score* 0.84. Model ini dipilih karena performanya yang unggul dibandingkan model lain dan telah didemonstrasikan kepada LPPM UMN untuk klasifikasi judul penelitian.

Kata kunci: Judul Penelitian, Klasifikasi Teks, *Logistic Regression*, LPPM UMN, TF-IDF



***Implementation of Logistic Regression Algorithm for Text Classification on
Research Titles Owned by Lecturers of Universitas Multimedia Nusantara***

Felix Ferdianto

ABSTRACT

Research and Community Service Institute of Multimedia Nusantara University (LPPM UMN) manages the grouping of UMN lecturer research titles manually, involving screening of more than 3000 research titles belonging to UMN lecturers from 2018-2023 without any additional systems. To increase efficiency, a text classification system was built using the logistic regression algorithm. This algorithm models the relationship between binary outcomes and predictor variables, using L1 penalty and a saga solver. This model utilizes the TF-IDF (Term Frequency Inverse Document Frequency) data preprocessing library, which is a library for calculating the frequency of importance of each word that appears in the dataset. This model also uses a dataset without going through a data cleaning process by using a dataset division of 80% for training and 20% for testing. The test results show that this model is the best model with an accuracy of 90.01%, precision 0.86, recall 0.82, and f1-score 0.84. This model was chosen because of its superior performance compared to other models and has been demonstrated to LPPM UMN for classification of research titles.

Keywords: *Logistic Regression, LPPM UMN, Research Title, Text Classification, TF-IDF*



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR GAMBAR	xi
DAFTAR GAMBAR	xi
DAFTAR TABEL	xii
DAFTAR KODE	xiii
DAFTAR LAMPIRAN	xiv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	3
1.3 Batasan Permasalahan	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	5
BAB 2 LANDASAN TEORI	6
2.1 Penelitian Terdahulu	6
2.2 Klasifikasi Teks	8
2.3 Logistic Regression	9
2.4 Data Cleaning	10
2.5 Count Vectorizer	15
2.6 TF-IDF	17
2.7 Word2Vec	20
2.8 Sustainable Development Goals (SDG)	22
BAB 3 METODOLOGI PENELITIAN	25
3.1 Metodologi Penelitian	25
3.2 Pengumpulan Data	26
3.3 Perancangan Model	28
BAB 4 HASIL DAN DISKUSI	34
4.1 Spesifikasi Sistem	34
4.2 Import Libraries	34
4.3 Load Dataset	35
4.4 Data Cleaning	36
4.5 Train Model	38
4.6 Evaluate Model	38
4.7 Uji Coba dan Evaluasi	40
4.7.1 Hasil Uji Coba	41
4.7.2 Evaluasi	48
4.8 Tanggapan LPPM UMN	54
BAB 5 SIMPULAN DAN SARAN	56
5.1 Simpulan	56
5.2 Saran	56



UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

Gambar 2.1	Contoh hasil <i>Count Vectorizer</i>	16
Gambar 2.2	Contoh hasil <i>Word2Vec</i>	21
Gambar 3.1	<i>Dataset</i> dari Hugging Face	27
Gambar 3.2	Validasi <i>Dataset</i>	27
Gambar 3.3	<i>Flowchart</i> Utama	28
Gambar 3.4	<i>Flowchart</i> Data Cleaning	29
Gambar 3.5	<i>Flowchart</i> Data Preprocessing	30
Gambar 3.6	<i>Flowchart</i> Train Model	31
Gambar 3.7	<i>Flowchart</i> Evaluate Model	32
Gambar 4.1	Tampilan <i>Dataset</i>	36
Gambar 4.2	Hasil data yang telah dibersihkan	37
Gambar 4.3	Hasil Evaluasi Model	39
Gambar 4.4	Grafik Hasil Uji Coba Skenario 1	44
Gambar 4.5	Grafik Tingkat Akurasi Skenario 1	44
Gambar 4.6	Grafik Hasil Uji Coba Skenario 2	47
Gambar 4.7	Grafik Tingkat Akurasi Skenario 2	47
Gambar 4.8	Hasil ROC dan AUC Pada Model dengan TF-IDF Dengan <i>Data Cleaning</i>	49
Gambar 4.9	Hasil ROC dan AUC Pada Model dengan TF-IDF Tanpa <i>Data Cleaning</i>	50
Gambar 4.10	Hasil <i>Learning Curve</i> Pada Model dengan TF-IDF Dengan <i>Data Cleaning</i>	51
Gambar 4.11	Hasil <i>Learning Curve</i> Pada Model dengan TF-IDF Tanpa <i>Data Cleaning</i>	53



DAFTAR TABEL

Tabel 2.1	Penelitian Terdahulu	6
Tabel 2.2	Contoh Data Judul Penelitian Sebelum dan Setelah <i>Case Folding</i>	11
Tabel 2.3	Contoh Data Judul Penelitian Sebelum dan Setelah Dihapus <i>Non-alphabet Character</i>	12
Tabel 2.4	Contoh Data Judul Penelitian Sebelum dan Setelah Dihapus <i>stopwords</i>	13
Tabel 2.5	Contoh Data Judul Penelitian Sebelum dan Setelah <i>Tokenization</i>	14
Tabel 2.6	Contoh Data Judul Penelitian Sebelum dan Setelah <i>Stemming</i>	15
Tabel 4.1	Hasil Uji Coba Skenario 1	43
Tabel 4.2	Hasil Uji Coba Skenario 2	46
Tabel 4.3	Hasil Uji Coba Semua Skenario	49



DAFTAR KODE

4.1	<i>Import Libraries</i>	35
4.2	Potongan kode <i>data cleaning</i>	36
4.3	Penggunaan <i>CountVectorizer</i>	38
4.4	Penggunaan model <i>logistic regression</i>	38
4.5	Evaluasi Model	39
4.6	Penggunaan <i>library Count Vectorizer</i> , TF-IDF, dan <i>Word2Vec</i> tanpa melakukan <i>data cleaning</i>	41



DAFTAR LAMPIRAN

Lampiran 1	Form Bimbingan Pembimbing 1	60
Lampiran 2	Form Bimbingan Pembimbing 2	61
Lampiran 3	Transkrip Wawancara 1	62
Lampiran 4	Transkrip Wawancara 2	64
Lampiran 5	Dokumentasi Wawancara 1	65
Lampiran 6	Dokumentasi Wawancara 2	65
Lampiran 7	Hasil Validiasi Klasifikasi Judul Dari LPPM UMN	66
Lampiran 8	Hasil Turnitin	73

