

**PENGGUNAAN EKSTRAKSI FITUR *TF-IDF* DAN *FASTTEXT*
PADA KLASIFIKASI SENTIMEN ULASAN LINKEDIN
DENGAN METODE *LOGISTIC REGRESSION***

SKRIPSI



Oleh :

NABILA SYA'BANI WARDANA

20081010114

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR
2024**

**PENGGUNAAN EKSTRAKSI FITUR *TF-IDF* DAN *FASTTEXT*
PADA KLASIFIKASI SENTIMEN ULASAN LINKEDIN
DENGAN METODE *LOGISTIC REGRESSION***

SKRIPSI

Diajukan Untuk Memenuhi Sebagai Persyaratan Dalam Menempuh Gelar Sarjana
Komputer Program Studi Informatika



Oleh :

NABILA SYA'BANI WARDANA

20081010114

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR
2024**

**LEMBAR PENGESAHAN
SKRIPSI**

**Judul : PENGGUNAAN EKSTRAKSI FITUR *TF-IDF* DAN *FASTTEXT*
PADA KLASIFIKASI SENTIMEN ULASAN LINKEDIN DENGAN
METODE *LOGISTIC REGRESSION***

Oleh : NABILA SYA'BANI WARDANA

NPM : 20081010114

**Telah Diseminarkan Dalam Ujian Skripsi Pada :
Hari Selasa, Tanggal 21 Mei 2024**

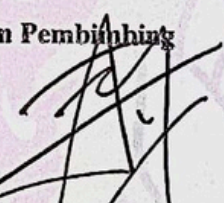
Mengetahui

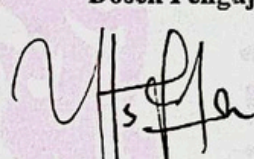
Dosen Pembimbing

Dosen Penguji

1.

1.


Firza Prima Aditiawan, S.Kom., MTI
NIP. 19860523/2021211 003


Yisti Vita Via, S.ST. M.Kom.
NIP. 19860425 2021212 001

2.

2.



Dr. Eng. Ir. Angraini Puspita Sari, ST., MT.
NIP. 222198 60 816400

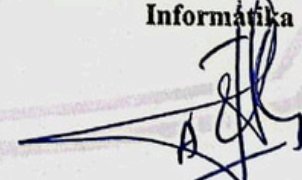

Agung Mustika Rizki, S.Kom., M.Kom.
NIP. 19930725 202203 1008

Menyetujui

**Dekan
Fakultas Ilmu Komputer**

**Koordinator Program Studi
Informatika**


Prof. Dr. Ir. Novirina Hendrasarie, M.T.
NIP : 196811261994032001


Fetty Tri Anggraeny, S.Kom. M.Kom
NIP. 19820211 2021212 005

SURAT PERNYATAAN ANTI PLAGIAT

Saya mahasiswa Program Studi Informatika Universitas Pembangunan Nasional “Veteran” Jawa Timur, yang bertanda tangan di bawah ini:

Nama : NABILA SYA'BANI WARDANA

NPM : 20081010114

Dengan ini menyatakan bahwa judul skripsi atau tugas akhir yang saya ajukan dan kerjakan, yang berjudul:

“PENGUNAAN EKSTRAKSI FITUR *TF-IDF* DAN *FASTTEXT* PADA
KLASIFIKASI SENTIMEN ULASAN LINKEDIN DENGAN METODE
LOGISTIC REGRESSION”


Bukan merupakan plagiat dari skripsi atau tugas akhir maupun penelitian orang lain dan juga bukan merupakan produk atau software yang saya beli dari pihak lain. Saya juga menyatakan bahwa skripsi ini adalah pekerjaan Saya sendiri, kecuali yang dinyatakan dalam daftar pustaka dan tidak pernah diajukan untuk syarat memperoleh gelar di Universitas Pembangunan Nasional “Veteran” Jawa Timur maupun institusi pendidikan lainnya.

Jika ternyata kemudian hari pernyataan ini terbukti tidak benar, maka Saya siap menerima segala konsekuensinya.

Surabaya, 3 Juni 2024

Hormat Saya,




5809BALX131331161
Nabila Sya'bani Wardana
NPM. 20081010114

PENGGUNAAN EKSTRAKSI FITUR *TF-IDF* DAN *FASTTEXT* PADA KLASIFIKASI SENTIMEN ULASAN LINKEDIN DENGAN METODE *LOGISTIC REGRESSION*

Oleh : NABILA SYA'BANI WARDANA

NPM : 20081010114

Program Studi : Informatika

Dosen Pembimbing : Firza Prima Aditiawan, S.Kom., MTI

Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.

ABSTRAK

Platform jejaring profesional seperti LinkedIn telah menjadi media sosial penting bagi individu untuk berinteraksi, berbagi informasi, dan membangun jaringan profesional. Meskipun LinkedIn telah memberikan manfaat signifikan bagi penggunanya, masih terdapat beberapa keterbatasan. Oleh karena itu, penting untuk memahami respons pengguna terhadap aplikasi ini. Penelitian sebelumnya menunjukkan bahwa analisis sentimen dapat menjadi alat yang efektif dalam memahami tanggapan dan respon pengguna. Penelitian ini akan menganalisis sentimen ulasan pengguna aplikasi LinkedIn menggunakan metode Regresi Logistik, dengan memperhitungkan penggunaan Ekstraksi Fitur *TF-IDF* dan Ekspansi Fitur *FastText*. Regresi Logistik dipilih karena efektif dalam menangani masalah klasifikasi sentimen biner dan memiliki kecepatan pelatihan yang relatif tinggi. Berdasarkan metrik evaluasi, dapat diamati bahwa klasifikasi sentimen ulasan pengguna LinkedIn dengan algoritma *Logistic Regression* berdasarkan pendekatan *TF-IDF* dan *FastText* menggunakan parameter hasil uji yang paling optimal mencapai akurasi sebesar 95,83%, presisi sebesar 96,54%, recall sebesar 96,38%, dan F1-Score sebesar 93,01%, sehingga dapat memberikan wawasan bagi pengembang LinkedIn untuk meningkatkan kualitas layanan.

Kata kunci: *LinkedIn, Analisis Sentimen, Regresi Logistik*

KATA PENGANTAR

Dengan penuh rasa syukur, penulis ingin mengucapkan terima kasih kepada Tuhan YME atas berkat dan rahmat-Nya yang telah memungkinkan penyelesaian skripsi berjudul " Penggunaan Ekstraksi Fitur *Tf-Idf* dan *Fasttext* pada Klasifikasi Sentimen Ulasan LinkedIn Dengan Metode *Logistic Regression* ".

Laporan skripsi ini disusun dalam rangka memenuhi persyaratan mata kuliah skripsi dan sebagai salah satu langkah penting dalam menyelesaikan Pendidikan tingkat Sarjana (S1) di Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur.

Penulis sadar akan adanya keterbatasan dalam penulisan skripsi ini, dan dengan rendah hati menerima saran serta kritik yang membangun. Semoga skripsi ini dapat memberikan manfaat bagi semua pihak, baik bagi pembaca maupun untuk penelitian mendatang.

Surabaya, 3 Juni 2024

Penulis,

Nabila Sya'bani Wardana

NPM. 20081010114

UCAPAN TERIMA KASIH

Penulisan laporan skripsi ini tidak dapat terwujud tanpa bantuan, motivasi, dan dukungan dari berbagai pihak. Oleh karena itu, dengan tulus, penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Akhmad Fauzi, M.MT., selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Ibu Prof. Dr. Ir. Novirina Hendrasarie, M.T., selaku Dekan Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur.
3. Ibu Fetty Tri Anggraeny, S.Kom., M.Kom., selaku Koordinator Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur dan Dosen Wali penulis.
4. Bapak Firza Prima Aditiawan, S.Kom., MTI selaku Dosen Pembimbing II yang meluangkan waktu, tenaga, serta pikiran untuk membimbing dan mengarahkan penulis selama proses penyelesaian skripsi.
5. Ibu Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT. selaku Dosen Pembimbing II yang telah memberikan arahan, dukungan, serta saran kepada penulis sehingga penulis dapat menyelesaikan penyusunan skripsi.
6. Seluruh Dosen Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur yang telah memberikan ilmunya kepada penulis selama perkuliahan.
7. Kedua orang tua penulis, Mbak Zahra, dan keluarga penulis yang senantiasa untuk mendoakan, dan memberikan dukuan selama proses penulisan skripsi ini.
8. Sahabat – sahabat penulis, Firda, Devina, Sebrina, Rafani, Mbak Farra, dan teman-teman di perkuliahan. Terimakasih karena selalu ada disamping penulis selama menyelesaikan perjalanan ini.

Penulis hanya bisa berharap, semoga Tuhan Yang Maha Esa senantiasa memberi perlindungan dan membalas semua kebaikan yang telah diberikan.

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI	i
ABSTRAK	iii
KATA PENGANTAR	iv
UCAPAN TERIMA KASIH.....	v
DAFTAR ISI.....	vi
DAFTAR TABEL.....	ix
DAFTAR GAMBAR	x
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	4
1.3. Tujuan.....	5
1.4. Manfaat.....	5
1.5. Batasan Masalah.....	5
BAB II TINJAUAN PUSTAKA.....	6
2.1. Penelitian Terdahulu.....	6
2.2. LinkedIn	8
2.3. Google Play Store.....	9
2.4. <i>Natural Language Processing (NLP)</i>	11
2.5. <i>Text Mining</i>	11
2.6. <i>Data Mining</i>	12
2.7. Analisis Sentimen.....	12
2.8. <i>Scraping</i>	13
2.9. <i>Preprocessing Text</i>	14
2.9.1. <i>Case Folding</i>	14
2.9.2. <i>Cleaning</i>	14
2.9.3. <i>Normalization</i>	14
2.9.4. <i>Stemming</i>	15
2.9.5. <i>Tokenizing</i>	15
2.10. Ekstraksi Fitur.....	15
2.11. <i>TF-IDF (Term Frequency-Inverse Document Frequency)</i>	15
2.12. <i>FastText</i>	16
2.15. <i>Logistic Regression</i>	19

2.16.	<i>Confusion Matrix</i>	22
BAB III METODOLOGI PENELITIAN.....		25
3.1.	Deskripsi Studi Kasus	25
3.2.	Alur Penelitian.....	26
3.3.	Studi Literatur.....	27
3.4.	Identifikasi Masalah	27
3.5.	Analisis dan Desain Sistem	27
3.5.1.	<i>Data Collection, Sorting, dan Labelling</i>	29
3.5.2.	<i>Text Preprocessing</i>	31
3.5.3.	Ekstraksi Fitur (<i>TF-IDF</i>)	33
3.5.4.	<i>Word Embedding (FastText)</i>	36
3.5.5.	<i>Data Splitting</i>	39
3.5.6.	Klasifikasi <i>Logistic Regression</i>	39
3.5.7.	Skenario Pengujian	44
3.5.8.	Evaluasi Performa.....	44
BAB IV HASIL DAN PEMBAHASAN		48
4.1.	Proses Akuisisi Data.....	48
4.2.	Pelabelan Data	50
4.3.	Eksplorasi Data.....	52
4.4.	<i>Preprocessing Text</i>	53
4.4.1.	Case Folding	53
4.4.2.	<i>Cleaning</i>	54
4.4.3.	Normalization.....	56
4.4.4.	Tokenizing.....	59
4.4.5.	Stemming	59
4.4.6.	Sortir Data	61
4.5.	<i>Word Cloud</i>	62
4.6.	Ekstraksi Fitur <i>TF-IDF</i>	63
4.7.	<i>Embedding Fasttext</i>	69
4.8.	Kombinasi Representasi Vektor	71
4.9.	Membuat Model Klasifikasi (<i>Logistic Regression</i>).....	73
4.10.	Pembagian Data	75
4.11.	Pelatihan dan Pengujian Model Klasifikasi	76
4.12.	Evaluasi Pelatihan dan Pengujian Metode (<i>Tuning Hyperparameter</i>).....	79

4.12.1. Split Data.....	79
4.12.2. Parameter Model Word Embedding (Fasttext)	83
4.12.3. Parameter Model Klasifikasi (<i>Logistic Regression</i>)	90
4.13. Perbandingan Penggunaan Representasi Vektor Fitur.....	98
4.14. Uji Hasil Prediksi pada Data Ulasan Asli.....	106
BAB V PENUTUP.....	109
5.1. Kesimpulan.....	109
5.2. Saran.....	109
DAFTAR PUSTAKA	111

DAFTAR TABEL

Tabel 3. 1 Contoh data hasil Scraping yang sudah dikonversi (.xlsx)	30
Tabel 3. 2 Contoh data hasil labelling	31
Tabel 3. 3 Contoh Data Hasil Penerapan Proses <i>Case Folding</i>	32
Tabel 3. 4 Contoh Data Hasil Penerapan Proses Cleaning.....	32
Tabel 3. 5 Contoh Data Hasil Penerapan Proses Normalization.....	32
Tabel 3. 6 Contoh Data Hasil Penerapan Tokenizing	33
Tabel 3. 7 Contoh Data Hasil Penerapan <i>Stemming</i>	33
Tabel 3. 8 Contoh korpus	34
Tabel 3. 9 Contoh perhitungan <i>TF</i>	34
Tabel 3. 10 Contoh perhitungan <i>DF</i>	34
Tabel 3. 11 Contoh perhitungan <i>IDF</i>	35
Tabel 3. 12 Contoh perhitungan <i>TF-IDF</i>	35
Tabel 3. 13 Contoh <i>matrix</i> hasil <i>TF-IDF</i>	36
Tabel 3. 14 Contoh representasi vektor <i>Fasttext</i> 2 dimensi	37
Tabel 3. 15 Contoh representasi vektor <i>Fasttext</i> 2 dimensi	38
Tabel 3. 16 Contoh representasi vektor gabungan	39
Tabel 3. 17 Contoh Data Ulasan Untuk Logistic Regression.....	41
Tabel 3. 18 Daftar set <i>split data</i>	44
Tabel 3. 19 Evaluasi metode pada model <i>word embedding Fasttext</i>	45
Tabel 3. 20 Evaluasi metode pada model klasifikasi <i>Logistic Regression</i>	46
Tabel 4.1 Distribusi data berdasarkan kategori sentimen.....	62
Tabel 4.2 hasil pengujian berdasarkan rasio pembagian data	82
Tabel 4.3 hasil pengujian berdasarkan ukuran vektor	86
Tabel 4.4 hasil pengujian berdasarkan nilai epoch.....	89
Tabel 4.5 hasil pengujian berdasarkan learning rate	93
Tabel 4.6 hasil pengujian berdasarkan jumlah iterasi	96
Tabel 4.7 parameter dengan hasil evaluasi terbaik.....	97
Tabel 4.8 hasil pengujian berdasarkan ekstraksi fitur	105

DAFTAR GAMBAR

Gambar 2. 1 Tampilan antarmuka halaman beranda aplikasi LinkedIn.....	9
Gambar 2. 2 Tampilan antarmuka halaman beranda <i>platform</i> Google Play Store	10
Gambar 2. 3 Ulasan salah satu aplikasi pada <i>platform</i> Google Play Store	10
Gambar 2. 4 Empat langkah utama teknik <i>NLP</i>	11
Gambar 2. 5 Ilustrasi analisis sentimen.....	13
Gambar 2. 6 Dataset hasil <i>scraping</i> ulasan aplikasi dari <i>platform</i> Google Play Store	13
Gambar 2. 7 Arsitektur <i>Word Embedding Fasttext</i>	17
Gambar 2. 8 Pendekatan <i>Machine Learning</i> dalam konteks analisis sentimen...	19
Gambar 2. 9 Arsitektur Klasifikasi <i>Logistic Regression</i>	20
Gambar 2. 10 Fungsi Sigmoid.....	22
Gambar 2. 11 <i>Confusion Matrix</i>	23
Gambar 3.1 Alur Penelitian	26
Gambar 3.2 Diagram Alur Sistem	28
Gambar 3.3 Flowchart Data Collection, Sorting, dan Labelling	29
Gambar 3.4 Diagram alur pra-pemrosesan data	31
Gambar 3.5 Diagram alur ekstraksi fitur <i>TF-IDF</i>	33
Gambar 3.6 Diagram alur <i>word embedding FastText</i>	36
Gambar 3.7 Diagram alur metodologi <i>logistic regression</i>	40
Gambar 4. 1 Id aplikasi linkedin	48
Gambar 4. 2 Hasil implementasi akuisisi data aplikasi LinkedIn	50
Gambar 4. 3 Hasil implementasi labelling data.....	51
Gambar 4. 4 Kamus <i>normalization data</i>	52
Gambar 4. 5 Hasil implementasi <i>Case Folding</i>	54
Gambar 4. 6 Hasil implementasi <i>Cleaning</i>	56
Gambar 4. 7 Hasil implementasi normalisasi	58
Gambar 4. 8 Hasil implementasi <i>Tokenizing</i>	59
Gambar 4. 9 Hasil implementasi <i>Stemming</i>	60
Gambar 4. 10 Visualisasi <i>Word Cloud</i>	62
Gambar 4. 11 Contoh perhitungan nilai <i>TF</i> pada data	64

Gambar 4. 12 Contoh hasil perhitungan nilai <i>DF</i>	65
Gambar 4. 13 Contoh hasil perhitungan nilai <i>IDF</i>	66
Gambar 4. 14 Contoh hasil perhitungan nilai <i>TF-IDF</i>	68
Gambar 4. 15 Ukuran matriks hasil representasi vektor <i>TF-IDF</i>	69
Gambar 4. 16 Contoh representasi <i>word embedding</i> pada kata ‘verifikasi’	70
Gambar 4. 17 Contoh representasi vektor berdasarkan kemiripan kata	71
Gambar 4. 18 Contoh representasi vektor fitur gabungan	72
Gambar 4. 19 Informasi ukuran dimensi dari pembagian data	76
Gambar 4. 20 <i>Output print</i> progres pelatihan setiap iterasi	77
Gambar 4. 21 <i>Output information</i> hasil pelatihan dan pengujian pertama	78
Gambar 4. 22 <i>Classification Report</i> hasil pengujian data test	78
Gambar 4. 23 Metrik Evaluasi klasifikasi rasio <i>split data</i> 70:30	80
Gambar 4. 24 Metrik Evaluasi klasifikasi rasio <i>split data</i> 80:20	81
Gambar 4. 25 Metrik Evaluasi klasifikasi rasio <i>split data</i> 90:10	82
Gambar 4. 26 <i>Classification Report</i> model <i>Fasttext</i> dengan vector size 200	84
Gambar 4. 27 <i>Classification Report</i> model <i>Fasttext</i> dengan vector size 300	85
Gambar 4. 28 <i>Classification Report</i> model <i>Fasttext</i> dengan nilai epochs 25	88
Gambar 4. 29 <i>Classification Report</i> model <i>Fasttext</i> dengan nilai epochs 50	89
Gambar 4. 30 <i>Classification Report</i> model <i>Logistic Regression</i> dengan $lr=0.03$ 91	
Gambar 4. 31 <i>Classification Report</i> model <i>Logistic Regression</i> dengan $lr=0.05$ 92	
Gambar 4. 32 <i>Classification Report</i> model <i>Logistic Regression</i> dengan num_iter=750	94
Gambar 4. 33 <i>Classification Report</i> model <i>Logistic Regression</i> dengan num_iter=1000	95
Gambar 4. 34 <i>Output information</i> hasil klasifikasi dengan vektor fitur gabungan	98
Gambar 4. 35 <i>Confusion Matrix</i> klasifikasi dengan vektor fitur gabungan	99
Gambar 4. 36 <i>Classification Report</i> hasil pelatihan dan pengujian klasifikasi dengan vektor <i>TF-IDF</i>	101
Gambar 4. 37 <i>Confusion Matrix</i> klasifikasi dengan vektor <i>TF-IDF</i>	102
Gambar 4. 38 <i>Classification Report</i> hasil pelatihan dan pengujian klasifikasi dengan vektor hasil model <i>Fasttext</i>	103

Gambar 4. 39 <i>Confusion Matrix</i> klasifikasi dengan vektor <i>FastText</i>	104
Gambar 4. 40 Contoh data uji 1 prediksi dan akutal	107
Gambar 4. 41 Contoh data uji 2 prediksi dan akutal	107
Gambar 4. 42 Contoh data uji 2 prediksi dan akutal	107