

**PENERAPAN *RANDOM FOREST* DENGAN ALGORITMA
LEVENSHTTEIN DISTANCE UNTUK ANALISIS SENTIMEN
TERHADAP ISU KEBIJAKAN BARU PROGRAM
BEASISWA LPDP**

Skripsi

**Disusun untuk memenuhi salah satu syarat
memperoleh gelar Sarjana Matematika**



**Dea Anggraeni
1305621035**

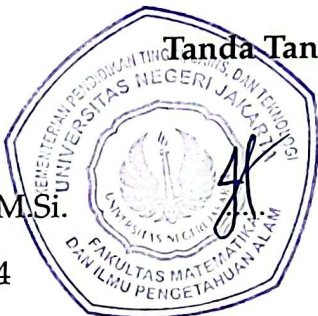
**PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI JAKARTA**

2026

LEMBAR PENGESAHAN UJIAN SKRIPSI
PENERAPAN RANDOM FOREST DENGAN ALGORITMA
LEVENSHTEIN DISTANCE UNTUK ANALISIS SENTIMEN TERHADAP
ISU KEBIJAKAN BARU PROGRAM BEASISWA LPDP

Nama : Dea Anggraeni

NIM : 1305621035

	Nama		Tanda Tangan	Tanggal
Penanggung Jawab				
Dekan :	Dr. Hadi Nasbey, S.Pd., M.Si.			26 / 01 / 2026
	NIP. 197909162005011004			
Wakil Penanggung Jawab				
Wakil Dekan I :	Dr. Meiliasari, S.Pd., M.Sc.			26 / 01 / 2026
	NIP. 197905042009122002			
Ketua :	Dr. Eti Dwi Wiraningsih, S.Pd., M.Si.			14 / 01 / 2026
	NIP. 198102032006042001			
Penguji Ahli :	Drs. Sudarwanto, M.Si., DEA			14 / 01 / 2026
	NIP. 196503251993031003			
Sekretaris :	Ibnu Hadi, M.Si.			14 / 01 / 2026
	NIP. 198107182008011017			
Pembimbing I :	Devi Eka Wardani M., S.Pd., M.Si.			19 / 01 / 2026
	NIP. 199005162019032014			
Pembimbing II :	Qorry Meidianingsih, S.Si., M.Si.			15 / 01 / 2026
	NIP. 199105192019032019			

Dinyatakan lulus ujian skripsi pada tanggal: 22 Desember 2025

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini, mahasiswa Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta:

Nama : Dea Anggraeni

No Registrasi : 1305621035

Program Studi : Matematika

Dengan ini menyatakan bahwa skripsi yang saya buat dengan judul *"Penerapan Random Forest dengan Algoritma Levenshtein Distance untuk Analisis Sentimen Terhadap Isu Kebijakan Baru Program Beasiswa LPDP"* adalah:

1. Dibuat sendiri, mengadopsi hasil kuliah, buku-buku, dan referensi acuan yang tertera di dalam referensi pada skripsi saya.
2. Bukan merupakan hasil duplikasi skripsi yang telah dipublikasikan atau pernah dipakai untuk mendapatkan gelar sarjana di Universitas lain kecuali pada bagian-bagian sumber informasi dicantumkan berdasarkan tata cara referensi yang semestinya.

Pernyataan ini dibuat dengan sesungguhnya dan saya bersedia menanggung segala akibat yang timbul jika pernyataan saya tidak benar.

Jakarta, 8 Desember 2025



Dea Anggraeni



KEMENTERIAN PENDIDIKAN TINGGI, SAINS DAN TEKNOLOGI
UNIVERSITAS NEGERI JAKARTA
PERPUSTAKAAN DAN KEARSIPAN

Jalan Rawamangun Muka Jakarta 13220

Telepon/Faksimili: 021-4894221

Laman: lib.unj.ac.id

**LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademika Universitas Negeri Jakarta, yang bertanda tangan di bawah ini, saya:

Nama : Dea Anggraeni
NIM : 1305621035
Fakultas/Prodi : FMIPA/Matematika
Alamat email : deaanggraeni2002@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan dan Kearsipan Universitas Negeri Jakarta, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah:

☒ Skripsi ☐ Tesis ☐ Disertasi ☐ Lain-lain (.....)

yang berjudul :

Penerapan Random Forest dengan Algoritma Levenshtein Distance untuk Analisis Sentimen
terhadap Isu Kebijakan Baru Program Beasiswa LPDP

Dengan Hak Bebas Royalti Non-Eksklusif ini Perpustakaan dan Kearsipan Universitas Negeri Jakarta berhak menyimpan, mengalihmediakan, mengelolanya dalam bentuk pangkalan data (*database*), mendistribusikannya, dan menampilkan/mempublikasikannya di internet atau media lain secara **fulltext** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan Universitas Negeri Jakarta, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 27 Januari 2026

Penulis

(Dea Anggraeni)
nama dan tanda tangan

ABSTRAK

DEA ANGGRAENI. Penerapan *Random Forest* dengan Algoritma *Levenshtein Distance* untuk Analisis Sentimen Terhadap Isu Kebijakan Baru Program Beasiswa LPDP. Skripsi, Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta. Desember 2025.

Media sosial menjadi wadah utama masyarakat dalam menyampaikan opini terhadap berbagai isu publik, termasuk kebijakan pemerintah. Kebijakan baru Program Beasiswa LPDP tahun 2024, yang memberikan kelonggaran bagi penerima untuk tidak wajib kembali ke Indonesia setelah studi di luar negeri, memunculkan beragam sentimen di *platform X (Twitter)*. Penelitian ini menganalisis sentimen publik terkait kebijakan tersebut dengan menerapkan algoritma *Levenshtein Distance* pada tahap normalisasi teks dan *Random Forest* sebagai model klasifikasi. Data dikumpulkan melalui *web scraping* terhadap unggahan berbahasa Indonesia di *platform X* periode 31 Oktober 2024 hingga 28 Februari 2025 menggunakan kata kunci 'LPDP'. Hasil evaluasi model pada tiga skenario menunjukkan peningkatan akurasi dari 83,58% pada *baseline* menjadi 84,05% pada model dengan normalisasi *Levenshtein Distance* murni, dan mencapai 86,82% pada model dengan normalisasi *Levenshtein Distance* berbasis *filtering rules* dan *scoring*. Analisis sentimen menggunakan model terbaik menghasilkan distribusi 26,7% positif, 34,3% negatif, dan 39,0% netral yang mencerminkan keberagaman persepsi masyarakat sekaligus memberikan gambaran komprehensif mengenai opini publik terhadap kebijakan baru LPDP.

Kata kunci: LPDP, Analisis Sentimen, Media Sosial, *Levenshtein Distance*, *Random Forest*

ABSTRACT

DEA ANGGRAENI. Application of Random Forest with the Levenshtein Distance Algorithm for Sentiment Analysis of the New LPDP Scholarship Program Policy. Mini Thesis, Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta. December 2025.

Social media has become a primary platform for the public to express opinions on various public issues, including government policies. The new 2024 LPDP Scholarship Program, which allows recipients not to return to Indonesia after studying abroad, has generated diverse sentiments on platform X (Twitter). This study analyzes public sentiment regarding the policy by applying the Levenshtein Distance algorithm for text normalization and Random Forest as the classification model. Data were collected through web scraping of Indonesian-language posts containing the keyword 'LPDP' from October 31, 2024, to February 28, 2025. Model evaluation across three scenarios showed an accuracy increase from 83.58% in the baseline scenario to 84.05% with pure Levenshtein Distance normalization, and up to 86.82% with normalization based on filtering rules and scoring. Sentiment analysis using the best performing model revealed a distribution of 26.7% positive, 34.3% negative, and 39.0% neutral, reflecting the diversity of public perception while providing a comprehensive overview of public opinion on the new LPDP policy.

Keywords: *LPDP, Sentiment Analysis, Social Media, Levenshtein Distance, Random Forest*

KATA PENGANTAR

Alhamdulillah rabbil 'alamin. Puji syukur penulis panjatkan kehadiran Allah SWT yang telah melimpahkan rahmat dan nikmat-Nya. Shalawat dan salam semoga selalu tercurahkan kepada Nabi Muhammad SAW, sehingga penulis senantiasa mendapatkan hidayah dan keberkahan dalam menuntut ilmu.

Penulisan skripsi berjudul “Penerapan *Random Forest* dengan Algoritma *Levenshtein Distance* untuk Analisis Sentimen Terhadap Isu Kebijakan Baru Program Beasiswa LPDP” ini diajukan guna memenuhi salah satu syarat untuk memperoleh gelar Sarjana Matematika pada Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta

Penulis menyadari bahwa keberhasilan penyelesaian skripsi ini tidak lepas dari bantuan, dukungan, dan bimbingan berbagai pihak. Dengan penuh hormat, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Hadi Nasbey, S.Pd., M.Si., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Jakarta beserta jajarannya.
2. Bapak Dr. Yudi Mahatma, M.Si., selaku Koordinator Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta.
3. Ibu Devi Eka Wardani Meganingtyas, S.Pd., M.Si. selaku dosen pembimbing I dan Ibu Qorry Meidianingsih, M.Si. selaku dosen pembimbing II yang telah memberikan arahan serta saran kepada penulis.

4. Bapak dan Ibu dosen yang telah membimbing dan mengajar penulis selama kurang lebih empat tahun, dengan ilmu yang sangat bermanfaat bagi penulis dalam penyelesaian skripsi ini serta sebagai bekal masa depan.
5. Diri sendiri, atas kesabaran, ketekunan, dan kerja keras dalam menyusun skripsi ini, serta kemampuan untuk tetap termotivasi menghadapi berbagai kesulitan yang muncul sepanjang prosesnya.
6. Kedua orang tua, Bapak Teguh Pamuji dan Ibu Sunarti, beserta seluruh keluarga, yang senantiasa memberikan doa, kasih sayang, motivasi, serta dukungan moral dan material, sehingga skripsi ini dapat terselesaikan dengan baik.
7. Teman-teman Program Studi Matematika terkhusus angkatan 2021 yang senantiasa menjadi teman bertumbuh serta tempat berbagi suka dan duka selama masa perkuliahan.
8. Semua pihak lain yang tidak dapat penulis sebutkan satu persatu, atas segala bentuk bantuan, dukungan, dan semangat yang telah diberikan kepada penulis.

Akhir kata, penulis memohon maaf apabila dalam penulisan skripsi ini terdapat kekurangan dan kesalahan, yang disebabkan oleh keterbatasan ilmu dan komunikasi. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun.

Semoga skripsi ini dapat memberikan manfaat bagi lingkungan akademik, khususnya Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, serta bagi pihak lain.

Jakarta, 8 Desember 2025



Dea Anggraeni

DAFTAR ISI

LEMBAR PENGESAHAN	i
LEMBAR PERNYATAAN	ii
ABSTRAK	iii
ABSTRACT	iv
KATA PENGANTAR	v
DAFTAR ISI	vii
DAFTAR TABEL	x
DAFTAR GAMBAR	xiii
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	5
1.3 Tujuan Penelitian	6
1.4 Batasan Masalah	6
1.5 Manfaat Penelitian	7
BAB 2 KAJIAN PUSTAKA	8
2.1 Analisis Sentimen	8
2.2 <i>Text Mining</i>	9
2.3 Teknik Pengambilan Data Teks	10
2.4 <i>Text Preprocessing</i>	12

2.5	<i>Algoritma Levenshtein Distance</i>	23
2.6	<i>Feature Extraction</i>	40
2.6.1	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	41
2.7	<i>Decision Tree</i>	47
2.8	Pengambilan Keputusan Kolektif	63
2.9	<i>Random Forest</i>	64
2.10	<i>K-Fold Cross Validation</i>	75
2.11	Evaluasi Kinerja Model Klasifikasi	77
BAB 3 METODOLOGI PENELITIAN		82
3.1	Data dan Sumber Data	82
3.2	Alur Penelitian	83
3.3	Diagram Alir	86
BAB 4 HASIL DAN PEMBAHASAN		87
4.1	<i>Scraping Data</i> dan <i>Filter Dokumen</i>	87
4.2	Pelabelan dan Validasi Dokumen	91
4.3	<i>Text Preprocessing</i>	95
4.3.1	<i>Data Cleaning</i>	95
4.3.2	<i>Case Folding</i>	97
4.3.3	<i>Tokenizing</i>	98
4.3.4	<i>Algoritma Levenshtein Distance</i>	99
4.3.5	<i>Stopword Removal</i>	107
4.3.6	<i>Stemming</i>	109
4.4	Pembagian Data Latih dan Data Uji	111
4.5	Pembobotan dengan <i>TF-IDF</i>	112
4.6	<i>10-Fold Cross Validation</i>	124
4.7	Fase Pelatihan Model <i>Random Forest</i>	127
4.8	Fase Prediksi Model <i>Random Forest</i>	138
4.8.1	Evaluasi Kinerja Model	138
4.8.2	Interpretasi Hasil Klasifikasi Sentimen	144

BAB 5 KESIMPULAN DAN SARAN	149
5.1 Kesimpulan	149
5.2 Saran	150
DAFTAR PUSTAKA	150
LAMPIRAN	156
RIWAYAT HIDUP	164



DAFTAR TABEL

Tabel 2.1	Contoh Proses <i>Data Cleaning</i>	13
Tabel 2.2	Contoh Proses <i>Case Folding</i>	15
Tabel 2.3	Contoh Proses <i>Tokenizing</i>	17
Tabel 2.4	Contoh Proses <i>Stopword Removal</i>	19
Tabel 2.5	Contoh Proses <i>Stemming</i>	22
Tabel 2.6	Perhitungan <i>Term Frequency</i>	44
Tabel 2.7	Perhitungan <i>Inverse Document Frequency</i>	45
Tabel 2.8	Perhitungan <i>Term Frequency-Inverse Document Frequency</i>	46
Tabel 2.9	Data Latih	53
Tabel 2.10	Nilai <i>TF-IDF</i> setiap Fitur pada Data Latih	54
Tabel 2.11	Nilai <i>TF-IDF</i> untuk Fitur 'anggaran'	56
Tabel 2.12	<i>Threshold</i> Optimal dan Δ Gini pada Setiap Fitur	59
Tabel 2.13	Data Uji	72
Tabel 2.14	Nilai <i>TF-IDF</i> setiap Fitur pada Data Uji	72
Tabel 2.15	<i>Confusion Matrix</i>	78
Tabel 2.16	Kriteria <i>Accuracy</i>	79
Tabel 4.1	Hasil Proses <i>Filtering</i> Data	90
Tabel 4.2	Hasil Proses <i>Labeling</i> Data	94
Tabel 4.3	Jumlah <i>Dokumen</i> untuk Setiap Kategori Sentimen	94
Tabel 4.4	Hasil Proses <i>Cleaning</i> Data	96
Tabel 4.5	Hasil Proses <i>Case Folding</i>	97
Tabel 4.6	Hasil Proses <i>Tokenizing</i>	99
Tabel 4.7	Hasil Normalisasi <i>Levenshtein Distance</i> tanpa <i>Filtering</i> <i>Rules dan Scoring</i>	100

Tabel 4.8	Hasil Normalisasi <i>Levenshtein Distance</i> dengan <i>Filtering Rules</i> dan <i>Scoring</i>	101
Tabel 4.9	Hasil Proses <i>Stopword Removal</i> (Tanpa Normalisasi <i>Levenshtein Distance</i>)	108
Tabel 4.10	Hasil Proses <i>Stopword Removal</i> (Normalisasi <i>Levenshtein Distance</i> Murni)	108
Tabel 4.11	Hasil Proses <i>Stopword Removal</i> (Normalisasi <i>Levenshtein Distance</i> dengan <i>Filtering Rules</i> dan <i>Scoring</i>)	109
Tabel 4.12	Hasil Proses <i>Stemming</i> (Tanpa Normalisasi <i>Levenshtein Distance</i>)	110
Tabel 4.13	Hasil Proses <i>Stemming</i> (Normalisasi <i>Levenshtein Distance</i> Murni)	110
Tabel 4.14	Hasil Proses <i>Stemming</i> (Normalisasi <i>Levenshtein Distance</i> dengan <i>Filtering Rules</i> dan <i>Scoring</i>)	111
Tabel 4.15	Hasil Pembagian Data Latih dan Data Uji	112
Tabel 4.16	Data Latih	113
Tabel 4.17	Perhitungan <i>Term Frequency</i> Data Latih	114
Tabel 4.18	Perhitungan <i>Inverse Document Frequency</i> Data Latih	115
Tabel 4.19	Perhitungan <i>TF-IDF</i> Data Latih	116
Tabel 4.20	Nilai <i>TF-IDF</i> Seluruh Fitur pada Data Latih	116
Tabel 4.21	Data Uji	118
Tabel 4.22	Perhitungan <i>Term Frequency</i> Data Uji	118
Tabel 4.23	Perhitungan <i>Inverse Document Frequency</i> Data Uji	120
Tabel 4.24	Perhitungan <i>TF-IDF</i> Data Uji	121
Tabel 4.25	Nilai <i>TF-IDF</i> Seluruh Fitur pada Data Uji	122
Tabel 4.26	Hasil <i>10-Fold Cross Validation</i> Uji Coba Ke-1	125
Tabel 4.27	Hasil <i>10-Fold Cross Validation</i> Uji Coba Ke-2	125
Tabel 4.28	Hasil <i>10-Fold Cross Validation</i> Uji Coba Ke-3	126
Tabel 4.29	<i>Dataset</i> Ilustrasi	128
Tabel 4.30	Perbandingan Parameter Model	130
Tabel 4.31	Nilai <i>TF-IDF</i> untuk Fitur Terpilih	133

Tabel 4.32	Hasil <i>Confusion Matrix</i> Uji Coba Ke-1	138
Tabel 4.33	Hasil <i>Confusion Matrix</i> Uji Coba Ke-2	140
Tabel 4.34	Hasil <i>Confusion Matrix</i> Uji Coba Ke-3	141
Tabel 4.35	Evaluasi Kinerja Model	142



DAFTAR GAMBAR

Gambar 2.1	Ilustrasi <i>Text Mining</i>	9
Gambar 2.2	Ilustrasi Penyisipan Karakter (<i>Insertion</i>)	24
Gambar 2.3	Ilustrasi Penghapusan Karakter (<i>Deletion</i>)	25
Gambar 2.4	Ilustrasi Pengubahan Karakter (<i>Subtitution</i>)	25
Gambar 2.5	Ilustrasi <i>Decision Tree</i>	48
Gambar 2.6	Ilustrasi <i>Split</i> di <i>Root Node</i> Berdasarkan Fitur 'anggaran'	61
Gambar 2.7	Ilustrasi <i>Split</i> Lanjutan Berdasarkan Fitur 'anak'	62
Gambar 2.8	Ilustrasi <i>Random Forest</i>	65
Gambar 2.9	Hasil Pembentukan <i>Decison Tree</i> φ_1	70
Gambar 2.10	Hasil Pembentukan <i>Decison Tree</i> φ_2	70
Gambar 2.11	Hasil Pembentukan <i>Decison Tree</i> φ_3	71
Gambar 2.12	Hasil Prediksi <i>Decison Tree</i> φ_1	73
Gambar 2.13	Hasil Prediksi <i>Decison Tree</i> φ_2	74
Gambar 2.14	Hasil Prediksi <i>Decison Tree</i> φ_3	74
Gambar 2.15	Ilustrasi <i>10-Fold Cross Validation</i>	77
Gambar 3.1	Diagram Alir Penelitian	86
Gambar 4.1	Hasil <i>Scraping Data</i> dengan <i>Keyword</i>	88
Gambar 4.2	Hasil <i>Scraping Data</i> dengan <i>Replies</i>	89
Gambar 4.3	Contoh Tampilan <i>Tweet</i> Sentimen Positif	92
Gambar 4.4	Contoh Tampilan <i>Tweet</i> Sentimen Negatif	93
Gambar 4.5	Contoh Tampilan <i>Tweet</i> Sentimen Netral	93
Gambar 4.6	Visualisasi φ_2 pada Program	132
Gambar 4.7	Kata Positif yang Sering Muncul	145
Gambar 4.8	Kata Negatif yang Sering Muncul	146
Gambar 4.9	Kata Netral yang Sering Muncul	147