

**DETEKSI INDIKASI GANGGUAN KESEHATAN MENTAL BERBASIS
TEKS MENGGUNAKAN NLP DENGAN TEKNIK
AUGMENTASI EDA**

SKRIPSI

Diajukan Untuk Memenuhi Sebagian Syarat Guna
Memperoleh Gelar Sarjana Komputer (S.Kom.)
Pada Program Studi Sistem Informasi



OLEH :

SHERLY DIAN TIARA

NPM: 2213030012

FAKULTAS TEKNIK DAN ILMU KOMPUTER (FTIK)
UNIVERSITAS NUSANTARA PERSATUAN GURU REPUBLIK INDONESIA
UN PGRI KEDIRI
2026

Proposal Skripsi oleh:

SHERLY DIAN TIARA

NPM: 2213030012

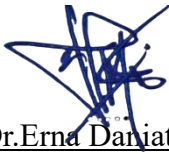
Judul:

**DETEKSI INDIKASI GANGGUAN KESEHATAN MENTAL BERBASIS TEKS
MENGUNAKAN NLP DENGAN TEKNIK
AUGMENTASI EDA**

Telah Diseminarkan dan Disetujui untuk Dilanjutkan
Guna Penulisan Skripsi / Tugas Akhir
Program Studi Sistem Informasi UN PGRI Kediri

Tanggal: 11 Juni 2026

Pembimbing I



Dr. Erna Daniati, M.Kom
NIDN. 0723058501

Pembimbing II



Arie Nugroho, M.Kom
NIDN. 0712108103

Skripsi oleh :

SHERLY DIAN TIARA

NPM: 2213030012

Judul

**DETEKSI INDIKASI GANGGUAN KESEHATAN MENTAL BERBASIS
TEKS MENGGUNAKAN NLP DENGAN TEKNIK
AUGMENTASI EDA**

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi

Program Studi Sistem Informasi FTIK UN PGRI Kediri

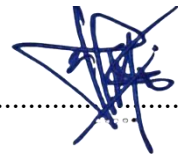
Tanggal: 25 Juni 2026

Dan Dinyatakan telah Memenuhi Persyaratan

Panitia Penguji :

1. Ketua : Dr. Erna Daniati, M.Kom

[.....]



2. Penguji 1 : Aidina Ristyawan, S.Kom, M.Kom

[.....]



3. Penguji 2 Arie Nugroho, S.Kom, M.M, M.Kom

[.....]



Mengetahui,

Dekan Fakultas Teknik dan Ilmu Komputer

Dr. Sulistiono, M.Si

NIDN. 0007076801

PERNYATAAN

Yang bertanda tangan di bawah ini saya,

Nama : Sherly Dian Tiara
Jenis Kelamin : Perempuan
Tempat/tgl. lahir : Nganjuk, 31 Juli 2003
NPM : 2213030012
Fak/Prodi. : FTIK/ S1-Sistem Informasi

menyatakan dengan sebenarnya, bahwa dalam Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak terdapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Kediri,

Yang Menyatakan



SHERLY DIAN TIARA

NPM: 2213030012

MOTTO

"Masa depan dibentuk oleh apa yang kita kerjakan hari ini,
bukan esok hari. Jadikan konsistensi sebagai pondasi, dan kedisiplinan sebagai
jalan menuju junya."

ABSTRAK

Sherly Dian Tiara: Deteksi Indikasi Gangguan Kesehatan Mental Berbasis Teks Menggunakan NLP dengan Teknik Augmentasi EDA, Skripsi, Sistem Informasi, Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri, 2026.

Kata Kunci: Deteksi Dini, Gangguan Mental, *Natural Language Processing*, *Easy Data Augmentation*, *Logistic Regression*.

Gangguan kesehatan mental merupakan isu serius yang memerlukan penanganan dini agar tidak berdampak fatal pada kualitas hidup penderitanya. Di era digital saat ini, masyarakat sering kali mengekspresikan kondisi emosional dan keluhan-keluh mereka secara terbuka melalui platform media sosial seperti Reddit. Fenomena ini membuka peluang untuk memanfaatkan teknologi komputasional dalam melakukan pemantauan awal terhadap kondisi psikologis pengguna melalui jejak digital berbasis teks. Namun, pemrosesan teks sering kali dihadapkan pada kendala ketidakseimbangan jumlah data antar kategori yang dapat menurunkan sensitivitas sistem dalam mendeteksi sinyal risiko pada kelas minoritas.

Penelitian ini bertujuan untuk mengolah dataset teks dari media sosial dan menerapkan teknik augmentasi data guna membangun model klasifikasi penyaringan awal (*screening*) gangguan mental. Metode penelitian dikembangkan mengacu pada kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang sistematis. Proses pra-proses melibatkan pembersihan teks, ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF), serta penerapan teknik *Easy Data Augmentation* (EDA) yang meliputi penggantian, penyisipan, pertukaran, dan penghapusan kata pada kelas minoritas. Algoritma *Logistic Regression* kemudian diimplementasikan sebagai model dasar (*baseline*) untuk membedakan enam kategori kondisi, yaitu *Anxiety*, *BPD*, *Bipolar*, *Depression*, *Schizophrenia*, dan *Mentalillness*.

Hasil pengujian menunjukkan bahwa penerapan teknik EDA memberikan dampak yang positif terhadap kemampuan pengenalan pola bahasa model pada kelas yang kekurangan data. Meskipun tingkat akurasi global mengalami sedikit penurunan dari 0.76 menjadi 0.74 akibat efek samping variasi data buatan, performa deteksi pada kategori *Mentalillness* mengalami peningkatan yang drastis. Nilai *Recall* pada kelas tersebut melonjak dari 0.28 menjadi 0.56, sementara nilai *F1-Score* meningkat dari 0.36 menjadi 0.59. Peningkatan metrik pada kelas minoritas ini membuktikan bahwa augmentasi teks berhasil memperkaya ekspresi linguistik, sehingga model menjadi lebih peka dan adil dalam mendeteksi indikasi pada seluruh kategori.

Kesimpulan dari penelitian ini menegaskan bahwa penggunaan teknik EDA relevan untuk menyeimbangkan sensitivitas sistem klasifikasi pada kasus penyaringan awal gangguan kesehatan mental. Kendati demikian, model komputasional ini memiliki keterbatasan karena hanya bersifat indikatif dan mutlak tidak dapat menggantikan kewenangan diagnosis klinis dari tenaga profesional medis. Evaluasi sistem juga telah divalidasi dengan melibatkan kepakaran psikolog profesional guna memastikan relevansi pola linguistik yang terdeteksi, **serta telah berhasil diimplementasikan ke dalam aplikasi *screening* deteksi awal berbasis web**. Untuk pengembangan selanjutnya, direkomendasikan agar penelitian mengimplementasikan dataset asli berbahasa Indonesia serta mengeksplorasi penggunaan algoritma *Deep Learning* tingkat lanjut guna memperoleh pemahaman konteks yang lebih mendalam.

KATA PENGANTAR

Puji Syukur Kami panjatkan kehadiran Allah Tuhan Yang Maha Kuasa, karena hanya atas perkenan dan rahmat-Nya, tugas penyusunan proposal penelitian ini dapat diselesaikan. Penyusunan proposal ini merupakan bagian dari rencana penelitian yang wajib dilakukan guna penyusunan skripsi, sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Teknik & Ilmu Komputer, Universitas Nusantara PGRI Kediri.

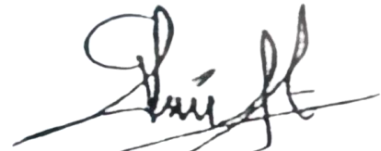
Pada kesempatan ini, penulis mengucapkan terima kasih dan penghargaan yang setulus-tulusnya kepada:

1. Dr. Zainal Afandi, M.Pd., selaku Rektor Universitas Nusantara PGRI Kediri yang selalu memberikan dorongan, motivasi, serta fasilitas terbaik kepada seluruh mahasiswa.
2. Dr. Sulistiono, M.Si selaku Dekan Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri, yang telah memberikan kemudahan birokrasi dan dukungan penuh selama masa studi.
3. Dr. Sucipto, M.Kom selaku Ketua Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri, atas segala arahan dan motivasinya kepada mahasiswa.
4. Ibu Erna Daniati M.Kom dan Arie Nugroho, M.Kom. selaku Dosen Pembimbing I dan Pembimbing II. Terima kasih yang tak terhingga atas kesabaran, waktu, tenaga, serta keikhlasannya dalam memberikan bimbingan, koreksi, dan wawasan yang sangat berharga hingga skripsi ini dapat terselesaikan dengan baik.
5. Seluruh Bapak dan Ibu Dosen Program Studi Sistem Informasi yang telah mentransfer ilmu, pengalaman, serta membekali penulis dengan wawasan teknologi yang sangat bermanfaat selama masa perkuliahan.
6. Kedua orang tua dan seluruh keluarga besar tercinta. Terima kasih atas setiap peluh, pengorbanan yang tak ternilai, serta doa-doa yang tidak pernah putus di setiap sujudnya, sehingga penulis selalu diberikan kelancaran dan kekuatan dalam menyelesaikan pendidikan ini.

7. Teman-teman seperjuangan Program Studi Sistem Informasi angkatan 2022 yang selalu menjadi tempat bertukar pikiran, berkeluh kesah, serta saling memberikan semangat dalam menyelesaikan skripsi ini.
8. Serta seluruh pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu, yang secara langsung maupun tidak langsung telah memberikan sumbangsih besar terhadap penyelesaian skripsi ini

Disadari bahwa proposal ini masih memiliki banyak kekurangan dan keterbatasan, baik dari segi penulisan maupun konten. Oleh karena itu, tegur sapa, kritik, dan saran-saran yang membangun dari berbagai pihak sangat diharapkan demi penyempurnaan proposal ini dan pelaksanaan penelitian di masa mendatang.

Kediri, 11 Juni 2026



SHERLY DIAN TIARA
NPM: 2213030012

DAFTAR ISI

Halaman Sampul	i
Halaman Persetujuan	ii
Halaman Pengesahan	iii
Halaman Pernyataan	iv
Halaman Motto	v
ABSTRAK	vi
KATA PENGANTAR	vii
DAFTAR ISI	ix
DAFTAR GAMBAR	xi
DAFTAR TABEL	xii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN	1
A. Latar Belakang	1
B. Batasan Masalah	4
C. Rumusan Masalah	4
D. Tujuan Penelitian	5
E. Manfaat Penelitian	5
BAB II KAJIAN PUSTAKA	7
A. Kajian Teori	7
B. Kajian Hasil Penelitian Terdahulu	9
C. Kerangka Berpikir	13
BAB III METODE PENELITIAN	15
A. Business Understanding	15
B. Data Understanding	16
C. Data Preparation	16
D. Modeling atau <i>Training</i>	17
E. Evaluation	18
F. Deployment	19
BAB IV HASIL DAN PEMBAHASAN	19
A. Data Hasil Penelitian	21
B. Pembahasan	25
BAB V PENUTUP	29

A.	Kesimpulan	29
B.	Saran	31
DAFTAR PUSTAKA		33
LAMPIRAN		37

DAFTAR GAMBAR

Gambar 2.1 Kerangka Berpikir	14
Gambar 3.1 Alur Penelitian	15
Gambar 3.2 <i>Confusion Matix</i>	18
Gambar 4.1 Distribusi Label	22
Gambar 4.2 Distribusi label setelah diaugmentasi	23
Gambar 4.3 <i>Confusion Matrix</i> model tanpa augmentasi EDA.....	24
Gambar 4.4 <i>Confusion Matrix</i> model dengan augmentasi EDA.....	24
Gambar 4.5 <i>Deployment</i> berbasis web	27

DAFTAR TABEL

Tabel 2.1 Kajian Terdahulu.....	9
Tabel 4.1 Perbandingan model.....	24
Tabel 4. 2 Validitas	27

DAFTAR LAMPIRAN

Lampiran 1. Surat Keterangan Validasi Pakar	37
Lampiran 2. Kartu Bimbingan.....	38
Lampiran 3. Surat Keterangan Bebas Similarity	39
Lampiran 4. Halaman Awal Hasil Cek Similarity.....	40
Lampiran 5. Bukti <i>Screensoot</i> Submit Artikel	40
Lampiran 6 Berita Acara Ujian Skripsi.....	41
Lampiran 7 Lembar Revisi Ujian Skripsi.....	42

BAB I

PENDAHULUAN

A. Latar Belakang

Gangguan kesehatan mental seperti depresi, kecemasan, dan gangguan terkait stres merupakan isu kesehatan masyarakat yang berdampak pada kualitas hidup, produktivitas, relasi sosial, serta meningkatkan risiko komorbiditas fisik. Di era digital, individu semakin sering mengekspresikan keluhan-kesah, emosi, dan pengalaman hidup melalui media sosial. Fenomena ini membuka peluang pemanfaatan pendekatan komputasional untuk pemantauan dan deteksi dini sinyal gangguan mental melalui teks yang dipublikasikan pengguna (Yusuf et al., 2023).

Riset internasional menunjukkan bahwa bahasa di media sosial dapat memuat indikator psikologis yang relevan untuk mengidentifikasi kondisi mental tertentu, termasuk depresi dan self-harm. Kajian sistematis di bidang ini juga menekankan pentingnya desain studi yang ketat, kualitas pelabelan, serta perhatian terhadap bias data dan validitas konstruk ketika memprediksi status kesehatan mental dari data media sosial (Situmorang & Purba, 2024).

Dalam ranah data dan benchmark, beberapa dataset berbasis Reddit telah banyak digunakan untuk penelitian deteksi gangguan mental. Contohnya RSDD (*Reddit Self-reported Depression Diagnosis*) yang memuat pengguna yang mengaku didiagnosis depresi beserta pengguna kontrol yang disesuaikan (Mohan et al., 2024). serta SMHD yang mencakup berbagai kondisi kesehatan mental dan memfasilitasi eksplorasi perbedaan penggunaan bahasa antar-kondisi. Selain itu, rangkaian evaluasi eRisk (CLEF) menyoroti tantangan *early risk prediction*, yaitu mendeteksi risiko sedini mungkin dari rangkaian posting yang bertambah seiring waktu (Mustofa et al., 2025).

Meskipun berharga, dataset berbahasa Inggris seperti Reddit memiliki keterbatasan ketika digunakan sebagai dasar untuk konteks Indonesia. Bahasa Indonesia di media sosial memiliki karakteristik khas misalnya slang, singkatan, campur kode, dan ragam daerah—sementara faktor sosio-kultural (norma, praktik, dan konteks lokal) dapat memengaruhi cara individu mengekspresikan distress. Sejalan dengan itu, penelitian pemodelan penalaran berbasis konteks budaya pada bahasa Indonesia menunjukkan bahwa penambahan konteks lokasi/budaya dapat meningkatkan kemampuan model memahami makna dan penalaran, sehingga aspek sosio-kultural patut dipertimbangkan dalam pemrosesan teks berbahasa Indonesia (Koto et al., 2024).

Di Indonesia, penelitian mengenai deteksi depresi atau kesehatan mental yang menggunakan NLP mulai meningkat (Zhang, 2021), contohnya deteksi depresi di Twitter/X dengan Word2Vec dan LSTM, serta studi deteksi awal gangguan mental berbasis sentimen menggunakan metode stacking pada data media sosial (Faruqziddan et al., 2025). Namun, tantangan umum yang sering dihadapi adalah kurangnya data yang relevan dengan konteks bahasa Indonesia (Riyadi, 2024), termasuk variasi bahasa informal dan dampak sosio-kultural, yang bisa memengaruhi kinerja serta generalisasi model jika tidak diperhatikan dengan baik (Yuliana et al., 2025).

Di samping itu, masalah lainnya adalah ketidakseimbangan kelas dan variasi dalam ekspresi bahasa (Utama et al., 2024). Dalam penelitian-penelitian text mining di Indonesia, berbagai teknik telah diterapkan untuk mengatasi tantangan ini, seperti oversampling seperti SMOTE untuk meningkatkan kinerja pada data yang tidak seimbang, serta berbagai model klasifikasi dan metode representasi (TF-IDF). Dalam aspek augmentasi teks, metode dasar seperti EDA (penggantian sinonim, sisipan acak, pertukaran acak, penghapusan acak) terbukti efektif dalam meningkatkan kinerja klasifikasi, khususnya pada dataset yang berukuran kecil (Nursahid et al., 2025).

Namun, terdapat kekurangan dalam penelitian karena augmentasi berbasis sinonim (EDA standar) sering kali menghasilkan variasi kalimat yang tidak alami dan belum mampu menangkap karakteristik unik media sosial

Indonesia, seperti penggunaan slang, referensi budaya lokal, serta pola ekspresi emosional yang spesifik secara sosial dan budaya. Sebagian besar studi yang ada masih berfokus pada pemanfaatan model kompleks tanpa menaruh perhatian yang cukup pada pengembangan dataset yang teraugmentasi secara kontekstual untuk mengatasi keterbatasan data berbahasa Indonesia. Sehubungan dengan itu, penelitian ini hadir untuk menjawab kekurangan tersebut dengan menyelidiki penerapan augmentasi EDA yang digabungkan dengan fokus pada elemen konteks agar dataset yang dihasilkan lebih mencerminkan kondisi yang sebenarnya dan dapat meningkatkan kinerja model dasar dalam mendeteksi tanda-tanda gangguan mental

Berdasarkan penjelasan di atas, studi ini menitikberatkan pada pengolahan dataset publik yang terdapat di Kaggle (berasal dari Reddit) untuk tugas deteksi atau klasifikasi tanda-tanda gangguan mental yang berbasis teks (Daniati et al., 2025). Penelitian melaksanakan tahapan pembersihan data, praproses teks, dan percobaan pemodelan NLP. Jika diperlukan, dilakukan penambahan data teks secara terbatas dengan menggunakan teknik *Easy Data Augmentation* (EDA) untuk membantu menangani variasi ungkapan dan ketidakseimbangan kelas. Penelitian ini memberikan sumbangan penting dalam beberapa aspek berikut.

Pengembangan Metodologi Praproses Teks Menyajikan panduan terstruktur tentang langkah-langkah pembersihan data (cleaning) dan normalisasi teks dari platform Reddit, mencakup penanganan nilai null serta entri yang tidak valid seperti teks yang sudah dihapus (Pradhana et al., 2025).

Optimalisasi Dataset lewat Augmentasi Menunjukkan efektivitas penerapan teknik EDA (Penggantian Sinonim, Penyisipan Acak, Pertukaran Acak, dan Penghapusan Acak) dalam meningkatkan variasi data latih tanpa perlu mengumpulkan data baru dari platform media sosial lainnya.

Pembuatan Model Dasar Menciptakan model klasifikasi awal dengan memanfaatkan kombinasi TF-IDF dan *Logistic Regression* yang dapat berfungsi sebagai acuan perbandingan untuk penelitian berikutnya yang menggunakan model yang lebih rumit.

Referensi Praktis untuk Penyaringan Kesehatan Mental Menyediakan dokumentasi langkah-langkah pengolahan data yang bisa menjadi pedoman

dalam menciptakan sistem klasifikasi sinyal potensi risiko kesehatan mental sebagai alat penyaringan awal yang tidak bersifat klinis.

Pondasi Riset Kontekstual Indonesia Menjadi acuan awal bagi peneliti lain untuk melakukan replikasi penelitian serupa menggunakan data asli berbahasa Indonesia dengan mempertimbangkan aspek sosio-kultural yang lebih mendetail

B. Batasan Masalah

1. Penelitian ini hanya membahas pengolahan dan pemanfaatan dataset publik dari Kaggle yang bersumber dari Reddit untuk kebutuhan deteksi dini indikasi gangguan kesehatan mental berbasis teks. Penelitian tidak menggunakan data medis maupun mengambil data tambahan dari platform media sosial lain.
2. Augmentasi dilakukan untuk menambah variasi teks hanya dari data yang sudah ada pada dataset Kaggle (Reddit), menggunakan teknik augmentasi sederhana seperti EDA (misalnya penggantian, penyisipan, pertukaran, atau penghapusan kata). Tidak ada pengambilan data tambahan dari media sosial lain.
3. Penilaian hasil penelitian menggunakan metrik evaluasi klasifikasi (akurasi, *precision*, *recall*, *F1-score*). Selain itu, dilakukan validasi kesesuaian indikasi klinis melalui uji pakar (*expert judgment*) oleh psikolog profesional untuk memastikan relevansi pola bahasa, namun sistem ini tetap diposisikan sebagai alat penyaringan (screening) awal, bukan pengganti diagnosis medis resmi.

C. Rumusan Masalah

1. Bagaimana proses pengolahan dataset publik seperti pembersihan data dan praproses teks untuk kebutuhan deteksi/klasifikasi indikasi gangguan kesehatan mental berbasis teks?
2. Bagaimana penerapan augmentasi teks EDA pada dataset tersebut untuk menambah variasi teks dan membantu mengatasi keterbatasan data (misalnya variasi ekspresi atau ketidakseimbangan kelas)?
3. Bagaimana perbandingan kualitas performa model baseline (jika digunakan) sebelum dan sesudah penerapan augmentasi EDA berdasarkan

metrik evaluasi klasifikasi standar?

D. Tujuan Penelitian

- 1 Mengolah dan menyiapkan dataset publik dari Kaggle (bersumber dari Reddit) menjadi data yang lebih terstruktur dan terdokumentasi untuk kebutuhan penelitian NLP, melalui proses pembersihan data dan praproses teks.
- 2 Menerapkan teknik augmentasi teks sederhana (EDA) pada data yang sama untuk menambah variasi teks dan membantu mengatasi keterbatasan data seperti variasi ekspresi dan ketidakseimbangan kelas, tanpa menambah sumber data baru.
- 3 Melakukan evaluasi terhadap kualitas data dan kinerja model baseline menggunakan metrik klasifikasi standar sebagai acuan bagi penelitian lanjutan terkait deteksi indikasi gangguan kesehatan mental berbasis teks.
- 4 Mengimplementasikan model klasifikasi terbaik yang dihasilkan ke dalam sebuah purwarupa (*prototype*) aplikasi berbasis web sederhana sebagai instrumen simulasi penyaringan awal.

E. Manfaat Penelitian

1 Manfaat Akademis

- a. Memberikan contoh pengolahan dan praproses dataset publik berbasis teks untuk kebutuhan penelitian NLP pada topik kesehatan mental.
- b. Menjadi referensi metodologis terkait pembersihan data, penanganan missing/teks terhapus, pemilihan fitur TF-IDF, serta penerapan augmentasi teks sederhana (EDA) pada studi klasifikasi teks.
- c. Mendorong penelitian lanjutan yang lebih komprehensif dan replikatif pada topik deteksi/klasifikasi indikasi gangguan kesehatan mental berbasis teks, termasuk pengujian pada data berbahasa Indonesia asli di masa mendatang.

2 Manfaat Praktis

- a. Menyediakan pipeline dan dokumentasi pengolahan data yang dapat

digunakan sebagai acuan untuk membangun sistem klasifikasi/pemantauan sinyal risiko kesehatan mental berbasis teks sebagai *screening*, bukan diagnosis.

- b. Membantu peneliti untuk pengembang memahami dampak augmentasi EDA dan isu seperti ketidakseimbangan kelas terhadap performa model klasifikasi teks.
- c. Memberikan kontribusi awal untuk meningkatkan pemahaman dan kewaspadaan tentang kesehatan mental melalui pendekatan komputasional atau berhubungan dengan komputer dan proses perhitungan, dengan menegaskan bahwa hasil prediksi hanya bersifat indikasi awal dan tidak menggantikan penilaian tenaga profesional

DAFTAR PUSTAKA

- Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). *Sentiment Classification : Review of Text Vectorization Methods : Bag of.* 4(1), 27–33. <https://doi.org/https://doi.org/10.56471/slujst.v4i.266>
- Aditiawan, F. P., Sari, A. P., Komputer, F. I., & Anyar, G. (2024). *VISA : Journal of Visions and Ideas VISA : Journal of Visions and Ideas.* 4(3), 1359–1370. <https://doi.org/DOI: 47467/visa.v4i3.2835> Logistic
- Akinode, A. O., Ayadi, O. E., Ezerioha, C. C., Chika, P., Akadiri, O. O., Adepoju, D. A., & Adepoju, A. G. (2025). *Machine Learning Approaches for Early Detection of Mental Health Disorders Using Wearable Devices and Big Data Analytics.* 6–23. <https://doi.org/DOI: https://doi.org/10.53771/ijbpsa.2025.10.2.0077>
- Alwuqaysi, B. (2025). Social Sciences & Humanities Open A systematic review and meta-analysis of social media ' s influence on mental health , family functioning , and visual content. *Social Sciences & Humanities Open*, 12(June), 101793. <https://doi.org/10.1016/j.ssaho.2025.101793>
- Azizah, A. N., Falach, M., Wisma, I., Prastya, D., Purwitasari, D., Korespondensi, P., & Daring, K. K. (2023). *EASY DATA AUGMENTATION UNTUK DATA YANG IMBALANCE PADA EASY DATA AUGMENTATION FOR IMBALANCED DATA OF ONLINE HEALTH.* 10(5). <https://doi.org/10.25126/jtiik.2023107082>
- Bankruptcy, P. C. (2023). *Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan.* 22(1), 207–214. <https://doi.org/10.33633/tc.v22i1.7527>
- Bimantara, M. D., & Zufria, I. (2024). *Text Mining Sentiment Analysis On Mobile Banking Application Reviews Using TF-IDF Method With Natural Language Processing Approach.* 5(1). <https://doi.org/https://doi.org/10.35877/454RI.jinav2772>
- Daniati, E., Wibawa, A. P., Sakti, W., & Irianto, G. (2025). *Building a Narrative Event Dataset from Andersen ' s Fairy Tales for Literary and Computational Analysis.* 5(3), 354–362. <https://doi.org/DOI: 10.52088/ijesty.v5i3.910>

- Faruqziddan, M., Daniati, E., & Muzaki, M. N. (2025). *Pendekatan BERT Dalam Analisis Sentimen Terhadap Kominfo Di Media Sosial X*. 4(2), 107–116. <https://doi.org/DOI:https://doi.org/10.59095/ijcsr.v4i2.207>
- Jadia, H. (2023). *COMPARATIVE ANALYSIS OF SENTIMENT ANALYSIS TECHNIQUES: SVM, LOGISTIC REGRESSION, AND TF-IDF FEATURE EXTRACTION*. 10, 1182–1187. <https://doi.org/DOI:https://www.doi.org/10.56726/IRJMETs45265>
- Koto, F., Mahendra, R., Aisyah, N., & Baldwin, T. (2024). IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Transactions of the Association for Computational Linguistics*, 12, 1703–1719. https://doi.org/10.1162/tacl_a_00726
- Kunci, K. (2025). *PERBANDINGAN KINERJA MODEL MACHINE LEARNING UNTUK SENTIMEN* *Diajukan : Diterima : Diterbitkan : 13(2)*, 80–88. <https://doi.org/DOI:https://doi.org/10.31294/jki.v13i2.11051>
- Method, S. (2025). *Deteksi Dini Gangguan Kesehatan Mental berdasarkan Sentimen menggunakan Metode Stacking Early Detection of Mental Health Disorders based on Sentiment using*. 14, 271–280. <https://doi.org/10.29103/sistemasi.v14i1.1396>
- Mohan, M., Ayyalasomayajula, T., Agarwal, A., & Khan, S. (2024). *Reddit social media text analysis for depression prediction : using logistic regression with enhanced term frequency-inverse document frequency features*. 14(5), 5998–6005. <https://doi.org/10.11591/ijece.v14i5.pp5998-6005>
- Mustofa, B. A., Laksito, W., & Saptomo, Y. (2025). *Use of Natural Language Processing in Social Media Text Analysis*. 4(2), 2–5. <https://doi.org/https://doi.org/10.59934/jaiea>
- Nugroho, K. S., Akbar, I., Suksmawati, A. N., Komputer, I., Brawijaya, U., Mada, U. G., Teknik, F., & Malang, U. W. (2021). *DETEKSI DEPRESI DAN KECEMASAN PENGGUNA TWITTER*. *Ciastech*, 287–296. <https://doi.org/10.48550/arXiv.2301.04521>
- Nursahid, W., Nugroho, B. I., & Syefudin, S. (2025). *Optimalisasi Preprocessing Data Menggunakan Pendekatan CRISP-DM untuk Meningkatkan Kualitas Klasifikasi Penyakit Jantung*. 4(3), 3621–3626.

<https://doi.org/10.48550/arXiv.2301.04521>

- Open, B., Model, A. I., Menggunakan, G.-T., & Whatsapp, M. (2024). *Kata Kunci : Chatbot, OpenAi , Model GTP-3.5 Turbo, IDS Digital College .* 785–793. [https://doi.org/DOI: https://doi.org/10.59095/ijcsr.v4i2.207](https://doi.org/DOI:https://doi.org/10.59095/ijcsr.v4i2.207)
- Parapar, J., Martín-rodilla, P., Losada, D. E., & Crestani, F. (2024). *Overview of eRisk at CLEF 2024 : Early Risk Prediction on the Internet (Extended Overview).* 0700. [https://doi.org/DOI: 10.1007/978-3-031-42448-9_24](https://doi.org/DOI:10.1007/978-3-031-42448-9_24)
- Pradhana, A. H., Daniati, E., & Najibulloh, M. (2025). *Penerapan Bi-LSTM Untuk Named Entity Recognition Pada Teks Bahasa Indonesia.* 4(2), 96–106. [https://doi.org/DOI: 10.59095/ijcsr.v4i2.208](https://doi.org/DOI:10.59095/ijcsr.v4i2.208).
- Rahma, I. A., Suadaa, L. H., Timur, J., & Korespondensi, P. (2023). *PENERAPAN TEXT AUGMENTATION UNTUK MENGATASI DATA YANG TIDAK SEIMBANG PADA KLASIFIKASI TEKS BERBAHASA INDONESIA STUDI KASUS : DETEKSI JUDUL CLICKBAIT DAN KOMENTAR HATE SPEECH APPLICATION OF TEXT AUGMENTATION TO HANDLING IMBALANCED DATA PROBLEM IN INDONESIAN TEXT CLASSIFICATION CASE STUDY : CLICKBAIT HEADLINE AND HATE SPEECH COMMENT.* 10(6), 1329–1340. <https://doi.org/10.25126/jtiik.2023107325>
- Riyadi, S. (2024). *Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (JIS) Pada Twitter Dengan Perbandingan Metode Naive Bayes Dan Support Vector Machine.* 5(3), 801–809. <https://doi.org/DOI:10.55338/saintek.v5i3.2790>
- Sari, A. A., & Permatasari, H. (2025). *Prediksi Risiko Kesehatan Mental Berdasarkan Pola Penggunaan Perangkat Digital Menggunakan Algoritma Logistic Regression.* November, 151–161. [https://doi.org/DOI : http://doi.org/10.55635/jic.v7i2](https://doi.org/DOI:http://doi.org/10.55635/jic.v7i2)
- Singgale, Y. A. (2024). *Sentiment Classification of The Capsule Hotel Guest Reviews using Cross-Industry Standard Process for Data Mining (CRISP-DM).* *Jurnal Media Informatika Budidarma*, 8(1), 632. <https://doi.org/10.30865/mib.v8i1.7329>
- Situmorang, G. F., & Purba, R. (2024). *Deteksi Potensi Depresi dari Unggahan*

- Media Sosial X Menggunakan Teknik NLP dan Model IndoBERT*. 6(2), 649–661. <https://doi.org/10.47065/bits.v6i2.5496>
- Tyas, F. A., Basir, A., & Ardhin, A. E. (2026). *PREDIKSI RESIKO PENGGUNAAN MEDIA SOSIAL TERHADAP KESEHATAN MENTAL MENGGUNAKAN EXPLORATORY DATA ANALYSIS (EDA) DAN CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM) PREDICTING THE IMPACT OF SOCIAL MEDIA USE ON MENTAL HEALTH THROUGH THE USE OF .* 13(2), 487–498. <https://doi.org/DOI:https://doi.org/10.25126/jtiik.2026133>
- Utama, H., Daniati, E., & Masruro, A. (2024). *WEAK SUPERVISION DENGAN PENDEKATAN LABELING FUNCTION UNTUK ANALISIS SENTIMEN*. 3(1), 49–57. [https://doi.org/DOI: 10.59095/ijcsr.v3i1.93](https://doi.org/DOI:10.59095/ijcsr.v3i1.93).
- Yuliana, D., Ningrum, A., Daniati, E., Muzaki, M. N., & Informasi, S. (2025). *Perbandingan Model BERT dan RNN-LSTM pada Analisis Sentimen Aplikasi BRI Mobile*. 4(2), 75–85. [https://doi.org/DOI: 10.59095/ijcsr.v4i2.206](https://doi.org/DOI:10.59095/ijcsr.v4i2.206).
- Yusuf, M., Psikologi, P. S., Muhammadiyah, U., Timur, K., Studi, P., Agama, P., & Malang, U. M. (2023). *Spiritualitas , Kecemasan , dan Stres selama Pandemi COVID-19*. 6(1), 47–54. [https://doi.org/DOI : 10.15575/jpib.v6i1.14145](https://doi.org/DOI:10.15575/jpib.v6i1.14145)
- Zhang, T. (2021). *Natural language processing applied to mental illness detection : a narrative review*. December, 1–13. <https://doi.org/10.1038/s41746-022-00589-7>