

## SURAT KETERANGAN BEBAS PLAGIARISME



**Universitas Nusantara PGRI Kediri**  
**UPT. PERPUSTAKAAN, PUBLIKASI DAN INOVASI**  
Alamat: Kampus 1, Jl. KH. Ahmad Dahlan No.76 Kota Kediri 64112  
Telp. (0354) 771576, (0354) 771503, (0354) 771495, Fax. (0354) 771576  
Website: <http://ppi.unpkediri.ac.id/> Email: [perpustakaan@unpkediri.ac.id](mailto:perpustakaan@unpkediri.ac.id)

## SURAT KETERANGAN BEBAS SIMILARITY

Ketua UPT Perpustakaan, Publikasi dan Inovasi Universitas Nusantara PGRI Kediri menerangkan bahwa mahasiswa dengan identitas berikut:

Nama Mahasiswa : Sherly Dian Tiara  
NPM : 2213030012  
Program Studi : S1-Sistem Informasi

Judul Karya Ilmiah:

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

Dinyatakan sudah memenuhi syarat batas maksimal 30% *similarity* sesuai dengan ketentuan yang berlaku pada setiap subbab naskah Laporan **Proposal Tugas Akhir/Skripsi/Tesis** yang disusun.  
Demikian Surat Keterangan ini kami berikan untuk dapat dipergunakan seperlunya.



Kediri, 25 Desember 2025  
Ka UPT PPI,



Dr. Abdul Aziz Hunaifi, M.A

# Turnitin LLCC

## 2213030012\_SHERLY DIAN TIARA.docx

 SI-08

---

### Document Details

**Submission ID**

trn:oid::2945:393806794

**Submission Date**

Jun 17, 2026, 9:44 AM GMT+7

**Download Date**

Jun 17, 2026, 9:47 AM GMT+7

**File Name**

2213030012\_SHERLY DIAN TIARA.docx.pdf

**File Size**

683.8 KB

**36 Pages**

**7,206 Words**

**46,039 Characters**

# 17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

## Filtered from the Report

- Bibliography

## Top Sources

- 11%  Internet sources
- 6%  Publications
- 13%  Submitted works (Student Papers)

## Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Top Sources

- 11% Internet sources
- 6% Publications
- 13% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

|    |                |                                           |     |
|----|----------------|-------------------------------------------|-----|
| 1  | Student papers | Universitas Khairun on 2026-05-18         | <1% |
| 2  | Internet       | dspace.uii.ac.id                          | <1% |
| 3  | Internet       | subset.id                                 | <1% |
| 4  | Student papers | Institut Teknologi Sumatera on 2026-05-18 | <1% |
| 5  | Internet       | cdn.juris.id                              | <1% |
| 6  | Internet       | doaj.org                                  | <1% |
| 7  | Internet       | scholar.archive.org                       | <1% |
| 8  | Internet       | jurnal.dharmawangsa.ac.id                 | <1% |
| 9  | Internet       | kc.umh.ac.id                              | <1% |
| 10 | Internet       | etheses.uin-malang.ac.id                  | <1% |
| 11 | Internet       | jutei.ukdw.ac.id                          | <1% |

|    |                |                                                                           |     |
|----|----------------|---------------------------------------------------------------------------|-----|
| 12 | Student papers | UNIVERSITAS BUDI LUHUR on 2025-07-26                                      | <1% |
| 13 | Student papers | Udayana University on 2025-12-14                                          | <1% |
| 14 | Student papers | Universitas Pancasila on 2025-08-25                                       | <1% |
| 15 | Internet       | repository.upi.edu                                                        | <1% |
| 16 | Student papers | Universitas Budi Luhur on 2026-01-24                                      | <1% |
| 17 | Student papers | Universitas Budi Luhur on 2026-01-31                                      | <1% |
| 18 | Student papers | Universitas Gadjah Mada on 2025-07-01                                     | <1% |
| 19 | Student papers | Universitas Jambi on 2026-03-10                                           | <1% |
| 20 | Internet       | ceur-ws.org                                                               | <1% |
| 21 | Internet       | ejurnal.seminar-id.com                                                    | <1% |
| 22 | Publication    | Abidin Abidin. "ANALISIS SENTIMEN ULASAN APLIKASI SIREKAP MENGGUNAKAN ... | <1% |
| 23 | Student papers | Universitas Sanata Dharma on 2026-06-10                                   | <1% |
| 24 | Internet       | jacis.pubmedia.id                                                         | <1% |
| 25 | Internet       | jurnal.amikom.ac.id                                                       | <1% |

|    |                |                                                                                    |     |
|----|----------------|------------------------------------------------------------------------------------|-----|
| 26 | Student papers | UIN Walisongo on 2024-09-28                                                        | <1% |
| 27 | Student papers | Universitas Tarumanagara on 2025-12-11                                             | <1% |
| 28 | Internet       | jurnal.bsi.ac.id                                                                   | <1% |
| 29 | Student papers | Sekolah Teknik Elektro & Informatika on 2026-01-09                                 | <1% |
| 30 | Student papers | Tilburg University on 2026-01-09                                                   | <1% |
| 31 | Student papers | University of Northumbria at Newcastle on 2025-08-26                               | <1% |
| 32 | Internet       | digilib.uinsby.ac.id                                                               | <1% |
| 33 | Internet       | eprints.umpo.ac.id                                                                 | <1% |
| 34 | Internet       | link.springer.com                                                                  | <1% |
| 35 | Internet       | www.scribd.com                                                                     | <1% |
| 36 | Publication    | Emil Fatra, Thiara Tri Funny Manguma, Mastura. "Analisis Sentimen Media Sosial ... | <1% |
| 37 | Student papers | Forum Perpustakaan Perguruan Tinggi Indonesia Jawa Timur II on 2025-08-20          | <1% |
| 38 | Student papers | UPN Veteran Yogyakarta on 2026-05-05                                               | <1% |
| 39 | Student papers | Universitas Muhammadiyah Purwokerto on 2025-01-13                                  | <1% |

|    |                |                                                                 |     |
|----|----------------|-----------------------------------------------------------------|-----|
| 40 | Internet       | ejournal.gunadarma.ac.id                                        | <1% |
| 41 | Student papers | IAI KAPD Jawa Timur on 2015-10-07                               | <1% |
| 42 | Student papers | Institut Bisnis dan Teknologi Indonesia (INSTIKI) on 2025-06-08 | <1% |
| 43 | Student papers | Perpustakaan on 2025-08-06                                      | <1% |
| 44 | Student papers | UPN Veteran Yogyakarta on 2026-02-13                            | <1% |
| 45 | Student papers | Universitas Islam Indonesia on 2025-06-24                       | <1% |
| 46 | Student papers | Universitas Pelita Harapan                                      | <1% |
| 47 | Internet       | ejurnal.ubharajaya.ac.id                                        | <1% |
| 48 | Internet       | eprints.poltektegal.ac.id                                       | <1% |
| 49 | Internet       | repository.ub.ac.id                                             | <1% |
| 50 | Internet       | www.neliti.com                                                  | <1% |
| 51 | Student papers | UPN Veteran Yogyakarta on 2026-03-02                            | <1% |
| 52 | Student papers | UPN Veteran Yogyakarta on 2026-04-22                            | <1% |
| 53 | Student papers | Universitas Gunadarma on 2025-09-24                             | <1% |

|    |                |                                                                                    |     |
|----|----------------|------------------------------------------------------------------------------------|-----|
| 54 | Student papers | Universitas Muhammadiyah Sukabumi on 2026-01-20                                    | <1% |
| 55 | Internet       | text-id.123dok.com                                                                 | <1% |
| 56 | Internet       | www.concreteconstruction.net                                                       | <1% |
| 57 | Internet       | www.djkn.kemenkeu.go.id                                                            | <1% |
| 58 | Student papers | Istanbul Bilgi University on 2025-06-21                                            | <1% |
| 59 | Publication    | Iwa Ovyawan Herlistiono, Sriyani Violina. "Model Prediksi Risiko Stroke Menggun... | <1% |
| 60 | Student papers | Universitas Bengkulu on 2025-10-17                                                 | <1% |
| 61 | Student papers | Universitas Gadjah Mada on 2024-01-24                                              | <1% |
| 62 | Student papers | Universitas Gunadarma on 2025-09-25                                                | <1% |
| 63 | Student papers | Universitas Putera Batam on 2025-07-25                                             | <1% |
| 64 | Student papers | Universitas Sumatera Utara on 2026-06-05                                           | <1% |
| 65 | Internet       | ejurnal.mikroskil.ac.id                                                            | <1% |
| 66 | Internet       | ejurnal.umri.ac.id                                                                 | <1% |
| 67 | Student papers | Christian University of Maranatha on 2025-01-11                                    | <1% |

|    |                |                                                                                      |     |
|----|----------------|--------------------------------------------------------------------------------------|-----|
| 68 | Publication    | Dhianciantyan Windydaca Brata Putri, Ni Putu Wintariani, Dewi Puspita Apsari, ...    | <1% |
| 69 | Publication    | Jeovanni Wong, Intan Sanu, Hafiz Irsyad. "Implementasi TF-IDF, Cosine Similarity,... | <1% |
| 70 | Student papers | Universitas Pancasila on 2026-03-04                                                  | <1% |
| 71 | Student papers | Universitas Putera Batam on 2026-01-29                                               | <1% |

19

## BAB I

### PENDAHULUAN

Bab ini menguraikan fondasi dasar pelaksanaan penelitian yang mencakup latar belakang masalah mengenai urgensi deteksi dini gangguan kesehatan mental melalui teks media sosial, batasan masalah untuk menjaga fokus penelitian, rumusan masalah, tujuan yang ingin dicapai, serta manfaat teoretis dan praktis yang diharapkan dari luaran penelitian ini.

43

#### A. Latar Belakang

39

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

66 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. 57 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.

8 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 (kualitas dataset atau performa model jika dibuat) hanya menggunakan ukuran evaluasi yang umum pada klasifikasi, seperti akurasi, precision, recall, dan F1-score, serta pengecekan apakah label dan data sudah cukup jelas serta mewakili kebutuhan penelitian. Penelitian ini tidak melakukan uji klinis dan tidak memvalidasi secara medis.

## **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.

## 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.

2

## BAB II

### KAJIAN PUSTAKA

64

Bab ini menyajikan landasan teoretis, telaah pustaka, dan kerangka pemikiran yang menjadi acuan dalam penyelesaian penelitian. Pemaparan materi difokuskan pada konsep gangguan mental, terminologi *Natural Language Processing* (NLP), metode representasi fitur TF-IDF, karakteristik algoritma *Logistic Regression*, serta implementasi teknik *Easy Data Augmentation* (EDA) untuk menangani ketidakseimbangan data. Bab ini juga dilengkapi dengan resume penelitian terdahulu sebagai pembanding posisi keaslian riset.

#### A. Kajian Teori

##### 1. Gangguan Mental dan Deteksi Dini

Deteksi dini gangguan mental merujuk pada upaya mengenali indikasi awal (early signals) dari kondisi psikologis tertentu sebelum berkembang menjadi lebih berat (Akinode et al., 2025). Dalam konteks komputasional, deteksi dini sering dipandang sebagai tugas klasifikasi/prediksi risiko berdasarkan jejak digital (misalnya posting media sosial), dengan tetap menegaskan bahwa hasil prediksi bukan diagnosis klinis dan harus diperlakukan secara etis (Situmorang & Purba, 2024).

12

##### 2. Natural Language Processing (NLP)

NLP adalah cabang kecerdasan buatan yang mempelajari bagaimana komputer memproses, memahami, dan menghasilkan bahasa manusia. Pada penelitian ini, NLP digunakan untuk memproses teks (pembersihan) dan mengekstraksi fitur/representasi untuk membangun dataset dan (opsional) model baseline (Mustofa et al., 2025).

24

##### 3. Representasi Teks Menggunakan TF-IDF

###### 1. Konsep TF-IDF

TF-IDF (*Term Frequency–Inverse Document Frequency*) adalah metode representasi teks yang memberikan bobot pada kata berdasarkan:

1

1) TF (*term frequency*): seberapa sering sebuah kata muncul dalam satu dokumen.

2) IDF (*inverse document frequency*): seberapa jarang kata tersebut

muncul di seluruh dokumen.

Secara umum, bobot TF-IDF tinggi untuk kata yang sering muncul dalam satu dokumen namun jarang muncul pada dokumen lain, sehingga dianggap lebih informatif untuk membedakan dokumen (Abubakar et al., 2022). Konsep ini banyak digunakan dalam information retrieval dan klasifikasi teks (Bimantara & Zufria, 2024).

#### 4. *Logistic Regression* untuk Klasifikasi Teks

##### 1. Konsep *Logistic Regression*

*Logistic Regression* adalah metode klasifikasi yang memodelkan probabilitas suatu kelas berdasarkan kombinasi linear fitur input (Kunci, 2025). Untuk klasifikasi biner, *Logistic Regression* menggunakan fungsi sigmoid untuk menghasilkan probabilitas kelas. Dalam konteks multi-kelas, *Logistic Regression* dapat diperluas sehingga dapat digunakan untuk klasifikasi dengan lebih dari dua label (Sari & Permatasari, 2025).

##### 2. *Logistic Regression* + TF-IDF

Pada klasifikasi teks, kombinasi TF-IDF + *Logistic Regression* sering dipakai sebagai baseline karena (Aditiawan et al., 2024):

- a) sederhana dan mudah dijelaskan,
- b) efisien untuk data berukuran besar,
- c) cukup kuat untuk banyak tugas klasifikasi teks,
- d) mendukung regularisasi untuk mengurangi overfitting.

##### 3. Augmentasi Data Teks Menggunakan EDA

Augmentasi data adalah teknik untuk menambah variasi data latih dengan membuat versi “baru” dari teks yang sudah ada, tanpa mengambil data dari sumber lain (Rahma et al., 2023). Salah satu pendekatan sederhana yang populer adalah EDA (*Easy Data Augmentation*), yang terdiri dari empat operasi:

- a. *Synonym Replacement*: mengganti beberapa kata dengan

sinonim.

- b. *Random Insertion*: menyisipkan kata secara acak.
- c. *Random Swap*: menukar posisi kata secara acak.
- d. *Random Deletion*: menghapus kata secara acak.

Penelitian yang dilakukan oleh (Azizah et al., 2023) menunjukkan bahwa EDA dapat meningkatkan performa klasifikasi, terutama ketika dataset terbatas, karena model melihat lebih banyak variasi bentuk kalimat selama pelatihan.

## B. Kajian Hasil Penelitian Terdahulu

Terdapat beberapa penelitian terdahulu yang di jadikan referensi dan pembanding dalam penelitian ini, sebagai berikut:

Tabel 2.1 Kajian Terdahulu

| Penulis & Tahun                     | Judul Penelitian                                                                          | Dataset               | Metode                                                 | Hasil / Kontribusi                                                                  |
|-------------------------------------|-------------------------------------------------------------------------------------------|-----------------------|--------------------------------------------------------|-------------------------------------------------------------------------------------|
| Faruqziddan, Daniati, dkk (2025)    | Pendekatan <i>BERT</i> Dalam Analisis Sentimen Terhadap Koinfo Di Media Sosial X          | Media Sosial X        | <i>BERT</i> ( <i>bert-base-uncased</i> ), <i>VADER</i> | Akurasi 84%, membuktikan <i>BERT</i> efektif memahami konteks opini publik.         |
| Akmal Pradhana, Daniati, dkk (2025) | Penerapan <i>Bi-LSTM</i> Untuk <i>Named Entity Recognition</i> Pada Teks Bahasa Indonesia | Korpus Teks Indonesia | <i>Bi-LSTM</i> , <i>BIO Tagging</i>                    | Akurasi 99%, sangat unggul dalam ekstraksi entitas pada teks Indonesia.             |
| Ningrum, Daniati, dkk (2025)        | Perbandingan Model <i>BERT</i> dan <i>RNN-LSTM</i> pada Analisis Sentimen BRI Mobile      | Google Play Store     | <i>BERT</i> vs <i>RNN-LSTM</i>                         | <i>BERT</i> unggul (73%) dan membuktikan tokenisasi ganda menurunkan kualitas data. |

|                             |                                                                                                                                                                                     |                          |                                             |                                                                            |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|---------------------------------------------|----------------------------------------------------------------------------|
| Erna Daniati, dkk (2025)    | <i>Building a Narrative Event Dataset from Andersen's Fairy Tales</i>                                                                                                               | Dongeng Andersen         | <i>Hybrid Annotation, NLP</i>               | Menghasilkan dataset terstruktur berisi 2.874 kejadian naratif.            |
| Erna Daniati & Utama (2023) | <i>Analisis Sentimen Dengan Pendekatan Ensemble Learning Dan Word Embedding Pada Twitter</i>                                                                                        | Twitter                  | <i>Word Embedding, Ensemble Learning</i>    | Memperkuat dasar penggunaan representasi teks dalam klasifikasi emosi.     |
| (Tyas et al., 2026)         | <i>PREDIKSI Resiko Penggunaan Media Sosial Terhadap Kesehatan Mental Menggunakan Exploratory Data Analysis (Eda) Dan Cross Industry Standard Process For Data Mining (Crisp-Dm)</i> | Dataset Klasifikasi Teks | Teknik <i>EDA</i> ( <i>SR, RI, RS, RD</i> ) | Augmentasi EDA secara signifikan meningkatkan performa pada dataset kecil. |
| Method, (2025)              | <i>Early Detection of Mental Health Disorders based on Sentiment using Stacking</i>                                                                                                 | Media Sosial             | <i>Stacking Classifier (RF, GB, LR)</i>     | Akurasi 95,66%, mengungguli model tunggal dalam deteksi kesehatan mental.  |
| Ayyalasomayajula (2024)     | <i>Reddit social media text analysis for depression</i>                                                                                                                             | Reddit                   | <i>Logistic Regression, TF-IDF</i>          | Peningkatan performa dengan penambahan fitur selang waktu                  |

|                        | <i>prediction</i>                                                                                                                 |                      |                                | <i>posting.</i>                                                          |
|------------------------|-----------------------------------------------------------------------------------------------------------------------------------|----------------------|--------------------------------|--------------------------------------------------------------------------|
| Inamdar et al. (2023)  | <i>Machine Learning Driven Mental Stress Detection on Reddit Posts</i>                                                            | Reddit               | <i>BoW, TF-IDF, BERT, ELMo</i> | Skor F1 0,76 dalam mengklasifikasi stres berbasis teks.                  |
| Nugroho et al., (2021) | Deteksi Depresi & Kecemasan Twitter Menggunakan <i>Bi-LSTM</i>                                                                    | Twitter              | <i>Bi-LSTM, SMOTE</i>          | Akurasi 94,12%, <i>Bi-LSTM</i> lebih baik dari <i>LSTM</i> standar.      |
| Widjayanto (2025)      | <i>Depression Detection using CNN and BiLSTM with BERT variations</i>                                                             | <i>Twitter</i>       | <i>Hybrid CNN-BiLSTM, BERT</i> | Akurasi 84,67%, menunjukkan kekuatan model hybrid.                       |
| Badry Ali M. (2025)    | <i>Use of Natural Language Processing in Social Media Text Analysis</i>                                                           | <i>Media Sosial</i>  | <i>NLP, Stacking</i>           | Mendukung penggunaan NLP untuk deteksi dini sinyal gangguan mental.      |
| (Alwuqaysi, 2025)      | <i>A systematic review and meta-analysis of social media's influence on mental health, family functioning, and visual content</i> | <i>Literatur</i>     | <i>Critical Review</i>         | Menekankan etika dan validitas konstruk dalam prediksi kesehatan mental. |
| (Zhang, 2021)          | <i>Natural language processing applied to mental illness detection: a</i>                                                         | <i>Reddit (RSDD)</i> | <i>Klasifikasi Teks</i>        | Menjadi acuan dataset internasional untuk riset depresi online.          |

|                          |                                                                                                               |                        |                               |                                                                        |
|--------------------------|---------------------------------------------------------------------------------------------------------------|------------------------|-------------------------------|------------------------------------------------------------------------|
|                          | <i>narrative review</i>                                                                                       |                        |                               |                                                                        |
| (Parapar et al., 2024)   | <i>Overview of eRisk at CLEF 2024: Early Risk Prediction on the Internet (Extended Overview)</i>              | <i>Posting Reddit</i>  | <i>Early Risk Prediction</i>  | Fokus pada deteksi risiko sedini mungkin melalui urutan waktu posting. |
| Koto et al. (2024)       | <i>IndoCulture: Exploring Geographically Influenced Cultural</i>                                              | <i>Teks Indonesia</i>  | <i>Konteks Sosio-Kultural</i> | Menekankan pentingnya konteks lokal dalam penalaran teks Indonesia.    |
| Bankruptcy, (2023)       | <i>Penerapan Metode Oversampling SMOTE</i>                                                                    | <i>Data Perusahaan</i> | <i>SMOTE, Random Forest</i>   | SMOTE efektif meningkatkan performa pada data yang tidak seimbang.     |
| (Azizah et al., 2023)    | <i>Easy Data Augmentation Untuk Data Yang Imbalance Pada Konsultasi Kesehatan Daring</i>                      | <i>Forum Kesehatan</i> | <i>Easy Data Augmentation</i> | EDA untuk menangani kelas data yang <i>imbalance</i> .                 |
| (Aditiawan et al., 2024) | <i>Logistic Regression Classification with TF-IDF and FastText for Sentiment Analysis of LinkedIn Reviews</i> | <i>Teks Umum</i>       | <i>Logistic Regression</i>    | Menjelaskan efisiensi regresi logistik untuk klasifikasi teks besar.   |
| Singgalen (2024)         | <i>Sentiment Classification using</i>                                                                         | <i>Ulasan Hotel</i>    | <i>CRISP-DM</i>               | Memberikan standarisasi alur penelitian data mining yang               |

|                            |                                                                        |                |                                |                                                                                      |
|----------------------------|------------------------------------------------------------------------|----------------|--------------------------------|--------------------------------------------------------------------------------------|
|                            | <i>CRISP-DM</i>                                                        |                |                                | sistematis.                                                                          |
| Zhang et al. (2022)        | <i>Natural language processing applied to mental illness detection</i> | Literatur      | <i>NLP Review</i>              | Memberikan klasifikasi metode NLP untuk berbagai kondisi psikologis.                 |
| Khurana et al. (2023)      | <i>NLP: State of the art, trends and challenges</i>                    | Literatur      | <i>Tren Deep Learning</i>      | Mengidentifikasi tantangan data terbatas dalam tugas klasifikasi teks.               |
| Arif (2023)                | Krisis Kesehatan Mental Remaja di Indonesia                            | Data Sosial    | Analisis Deskriptif            | Memberikan justifikasi sosial mengenai urgensi deteksi dini.                         |
| Assegaf et al. (2024)      | Chatbot WhatsApp GPT-3.5 Kesehatan Mental                              | Chat Log       | GPT-3.5 (LLM)                  | Menunjukkan potensi kecerdasan buatan dalam merespons keluhan mental.                |
| Ibrahim et al. (2024)      | Deteksi Potensi Depresi Model IndoBERT                                 | Teks Indonesia | <i>Transformer (IndoBERT)</i>  | Model lokal Indonesia memberikan akurasi yang lebih spesifik konteks.                |
| Hamka & Yusuf (2023)       | <i>Spiritualitas, Kecemasan, dan Stres Pandemi</i>                     | Kuesioner      | Analisis Faktor Psikologis     | Menjelaskan hubungan antara kosa kata emosional dengan kondisi mental.               |
| Erna Daniati (2024)        | Dukungan Keputusan Pembelian dengan Analisis Sentimen                  | Media Sosial   | Analisis Sentimen, <i>MCDM</i> | Mengintegrasikan sentimen pengguna media sosial ke dalam sistem pendukung keputusan. |
| Utama, Daniati, dkk (2024) | <i>Weak Supervision</i>                                                | Twitter        | <i>Weak Supervision</i>        | Mengatasi kendala pelabelan                                                          |

|  |                                     |  |  |                                            |
|--|-------------------------------------|--|--|--------------------------------------------|
|  | dengan Pendekatan Labeling Function |  |  | manual pada data media sosial skala besar. |
|--|-------------------------------------|--|--|--------------------------------------------|

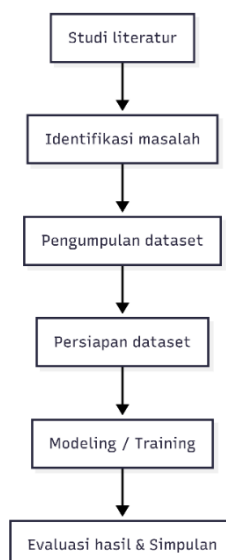
### C. Kerangka Berpikir

Penelitian ini berangkat dari kebutuhan ketersediaan dataset deteksi dini gangguan mental yang relevan untuk konteks pengguna Indonesia. Dataset internasional (misalnya RSDD/SMHD) telah mendukung riset deteksi depresi dan kondisi mental lain, namun perbedaan bahasa serta konteks sosio-kultural dapat menurunkan validitas dan generalisasi apabila langsung diadopsi (Jadia, 2023).

Penelitian ini memfokuskan pada pengembangan dataset berbahasa Indonesia yang terkurasi dan terdokumentasi. Setelah proses pengumpulan data, dilakukan pembersihan serta normalisasi bahasa informal. Selanjutnya, dilakukan pelabelan sesuai definisi operasional kelas, disertai pemeriksaan kualitas label dan distribusi kelas.

Untuk meningkatkan keragaman ekspresi dan mengurangi kekurangan data, diterapkan augmentasi. Selain augmentasi umum (misalnya EDA), penelitian ini menambahkan augmentasi konteks sosio-kultural agar dataset lebih representatif (Koto et al., 2024). Bila diperlukan, ketidakseimbangan kelas ditangani dengan strategi seperti SMOTE.

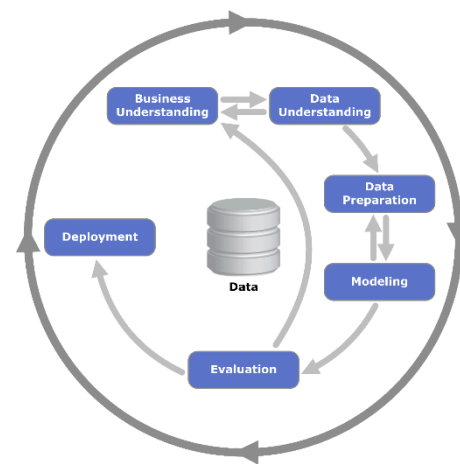
Output utama penelitian adalah dataset deteksi dini gangguan mental berbahasa Indonesia beserta dokumentasi proses, dan (opsional) baseline model untuk menunjukkan kegunaan dataset. Kerangka berpikir penelitian digambarkan pada Gambar 1.



### BAB III

## METODE PENELITIAN

Penelitian ini menggunakan kerangka CRISP-DM (Cross-Industry Standard Process for Data Mining) karena tahapannya jelas dan cocok untuk penelitian klasifikasi teks berbasis data. Tahapan CRISP-DM terdiri dari: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* (Singgalen, 2024). Alur penelitian digambarkan dalam diagram berikut:



#### A. Business Understanding

Permasalahan yang diangkat adalah bahwa banyak individu mengekspresikan kondisi emosional, keluh-kesah, dan pengalaman psikologis melalui tulisan, sehingga muncul peluang untuk memanfaatkan pendekatan komputasional guna mengklasifikasikan indikasi atau kategori gangguan kesehatan mental berdasarkan teks. Penelitian ini menggunakan dataset publik dari Kaggle yang bersumber dari Reddit, sehingga fokus penelitian diarahkan pada pengolahan data teks dan eksperimen pemodelan klasifikasi, bukan untuk melakukan diagnosis klinis. Tujuan pada tahap ini adalah menyusun alur pemrosesan teks untuk klasifikasi indikasi gangguan kesehatan mental berbasis teks, serta menguji dampak penerapan augmentasi teks EDA terhadap performa model baseline. Luaran dari tahap Business Understanding mencakup rumusan masalah dan batasan penelitian yang menegaskan bahwa hasil bersifat indikatif, serta penetapan target keluaran

sistem berupa prediksi label (misalnya *subreddit*) dan evaluasi menggunakan metrik klasifikasi standar.

### B. Data Understanding

Pada tahap *Data Understanding*, penelitian ini bertujuan untuk memahami isi dataset serta kondisi kualitas datanya sebelum diproses lebih lanjut. Data yang digunakan merupakan dataset publik dari Kaggle yang bersumber dari Reddit, dengan kolom utama yang digunakan meliputi (*title*, *selftext*, *created\_utc*, *over\_18*, dan *subreddit*). Pada tahap ini dilakukan pemeriksaan terhadap jumlah data dan tipe data tiap kolom, serta identifikasi nilai hilang (*missing value*) dan entri yang tidak valid seperti teks (*removed*) atau (*deleted*) yang dapat mengurangi informasi pada konten posting. Selain itu, penelitian menganalisis distribusi label subreddit untuk melihat apakah terjadi ketidakseimbangan kelas (*class imbalance*) antara kelas mayoritas dan minoritas. Variasi panjang teks juga diperiksa, karena perbedaan antara judul yang singkat dan isi posting yang panjang atau kosong dapat memengaruhi kualitas representasi teks. Luaran dari tahap Data Understanding adalah ringkasan kondisi data (misalnya jumlah missing, teks terhapus, dan jumlah kelas), serta statistik distribusi label dan potensi masalah yang perlu ditangani pada tahap persiapan data, seperti noise dan ketidakseimbangan kelas.

### C. Data Preparation

Pada tahap *Data Preparation*, penelitian ini bertujuan menyiapkan data agar siap digunakan oleh model klasifikasi :

- b) pembersihan data (*cleaning*), yaitu menangani nilai null pada kolom *selftext* (misalnya diganti menjadi string kosong), menormalisasi atau mengosongkan teks yang tidak valid seperti (*removed*) , serta menghapus URL, spasi berlebih, dan karakter yang tidak diperlukan.
- c) Membuat kolom teks utama, dengan menggabungkan *title* dan *selftext* menjadi satu kolom text menggunakan *format text = title + " " + selftext*, serta dapat ditambahkan normalisasi huruf seperti *lowercase* jika diperlukan.
- d) Filtering data, Hapus baris yang text-nya kosong/terlalu pendek (misalnya < 3–5 kata), karena kurang informatif.

- e) *Split* dataset, dataset dibagi menjadi data latih, validasi, dan uji (*train/validation/test*), dan disarankan menggunakan stratified split agar proporsi label *subreddit* tetap seimbang di setiap bagian data.
- f) Augmentasi teks EDA, pada data latih—terutama pada kelas minor—untuk menambah variasi teks, menggunakan teknik *synonym replacement*, *random insertion*, *random swap*, dan *random deletion* dengan aturan agar perubahan tidak merusak makna secara berlebihan.

#### D. Modeling atau *Training*

Pada tahap Modeling, tujuan penelitian adalah membangun model klasifikasi teks untuk memprediksi label *subreddit* berdasarkan isi teks.

- a. Komponen model baseline
  - 1) TF-IDF digunakan untuk mengubah teks menjadi fitur numerik (angka) agar bisa diproses oleh model.
  - 2) *Logistic Regression* digunakan sebagai algoritma klasifikasi untuk memprediksi label/kategori
- b. Alur pemodelan
  - 1) Input data Menggunakan kolom *text* sebagai masukan utama.
  - 2) Ekstraksi fitur Mengubah text menjadi vektor angka menggunakan TF-IDF.
  - 3) Pelatihan model Melatih Logistic Regression menggunakan vektor TF-IDF agar model mempelajari pola tiap label.
  - 4) Output prediksi Model menghasilkan prediksi label Model juga menghasilkan tingkat keyakinan atau akurasi prediksi.
- c. Eksperimen utama
  - 1) Model 1 TF-IDF dan Logistic Regression tanpa augmentasi tanpa EDA.
  - 2) Model 2 dengan EDA TF-IDF + Logistic Regression menggunakan data latih yang sudah ditambah variasinya dengan EDA.

#### E. Evaluation

Model yang telah dilatih dievaluasi menggunakan metrik standar deteksi objek, nilai prediksi dan nilai aktual. Seperti pada gambar berikut.

|     |         | Nilai Valid         |                     |
|-----|---------|---------------------|---------------------|
|     |         | Positif             | Negatif             |
| isi | positif | True Positive (TP)  | False Positive (FP) |
|     | negatif | False Negative (FN) | True Negative (TN)  |

Arti tiap komponen: *Confusion Matix*.

**TP:** model memprediksi *positif* dan memang *positif* (benar)

**TN:** model memprediksi *negatif* dan memang *negatif* (benar)

**FP:** model memprediksi *positif*, padahal sebenarnya *negatif* (salah)

**FN:** model memprediksi *negatif*, padahal sebenarnya *positif* (salah)

#### a) Precision

Precision merupakan metrik yang mengukur tingkat ketepatan prediksi positif yang dihasilkan model. Precision dihitung berdasarkan proporsi *True Positive* (TP) terhadap jumlah keseluruhan prediksi positif, termasuk *False Positive* (FP). Nilai precision yang tinggi menunjukkan bahwa model mampu melakukan deteksi objek secara tepat dengan jumlah kesalahan deteksi (FP) yang rendah.

$$Precision = \frac{TP}{TP+FP} \quad (3.1)$$

#### b) Recall

Recall mengukur sejauh mana model mampu mendeteksi seluruh objek yang seharusnya terdeteksi. Recall merupakan rasio antara *True Positive* (TP) dan jumlah total objek aktual yang terdiri dari *True Positive* dan *False Negative* (FN). Nilai recall yang tinggi menunjukkan bahwa model memiliki kemampuan deteksi yang baik dan mampu mengurangi risiko objek terlewatkan.

$$Recall = \frac{TP}{TP+FN} \quad (3.2)$$

#### c) Accuracy (Akurasi)

Accuracy (akurasi) merupakan metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan model klasifikasi secara keseluruhan, yaitu seberapa besar proporsi prediksi yang dilakukan oleh model sesuai dengan label sebenarnya. Dengan kata lain, accuracy menunjukkan persentase data yang berhasil diklasifikasikan dengan benar oleh model

terhadap seluruh data pengujian.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.3)$$

d) F1-Score

F1-score merupakan metrik evaluasi yang digunakan untuk mengukur keseimbangan antara precision dan recall dalam suatu model klasifikasi.

F1-score sangat berguna ketika dataset yang digunakan tidak seimbang atau ketika kesalahan prediksi pada kelas tertentu memiliki dampak yang signifikan.

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.4)$$

## F. Deployment

Penelitian ini berfokus pada pengembangan model, sehingga tahap deployment dilakukan dalam bentuk simulasi. Hasil model terbaik yang dihasilkan kemudian diekspor dan disimpan. Selanjutnya, dilakukan uji coba inferensi pada lingkungan Google Colab untuk memastikan model dapat dijalankan kembali dan menghasilkan prediksi secara konsisten. Selain itu, dikembangkan aplikasi web sederhana berbasis Streamlit sebagai media demonstrasi untuk memfasilitasi pengujian model, di mana pengguna dapat memasukkan teks dan sistem menampilkan hasil prediksi yang dihasilkan model (Open et al., 2024).



## BAB IV

### HASIL DAN PEMBAHASAN

Bab ini menyajikan seluruh data yang diperoleh dari hasil eksperimen komputasional serta analisis mendalam mengenai temuan riset tersebut. Pembahasan difokuskan pada paparan hasil eksplorasi dataset Reddit, proses transformasi dan pembersihan data, visualisasi perbandingan performa model *baseline* sebelum dan sesudah penerapan teknik *augmentasi* EDA, serta pembuktian validitas sistem baik secara statistik maupun berdasarkan kepakaran klinis psikolog.

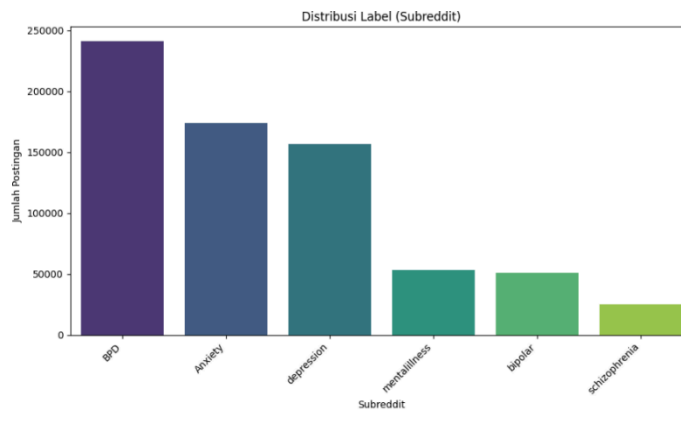
#### A. Data Hasil Penelitian

##### 1. Hasil Eksplorasi Data

Tahap awal penelitian dimulai dengan melakukan eksplorasi mendalam terhadap dataset yang telah dikumpulkan untuk memahami karakteristik awal data indikasi gangguan kesehatan mental dari Reddit. Peneliti melakukan pemeriksaan terhadap struktur dataset yang terdiri dari kolom *title*, *selftext*, dan *subreddit* sebagai label utama klasifikasi. Berdasarkan observasi awal, ditemukan bahwa sebagian besar data memiliki variasi panjang teks yang sangat beragam antara judul postingan dengan isi kontennya. Selain itu, identifikasi awal menunjukkan adanya masalah kualitas data berupa nilai *null* dan entri teks yang mengandung tanda *removed* atau *deleted* yang harus ditangani pada tahap berikutnya.

Selain pemeriksaan kualitas konten, juga dilakukan analisis terhadap distribusi label untuk mengetahui keseimbangan kelas dalam dataset. Hasil analisis statistik menunjukkan adanya ketidakseimbangan kelas (*class imbalance*) yang signifikan di antara kategori *subreddit*. Kondisi ini tergambar jelas pada Gambar 4.1, di mana kategori BPD memiliki jumlah sampel yang dominan melebihi 200.000 data, sementara kategori lain, seperti *mentalillness* dan *schizophrenia*, memiliki jumlah sampel yang jauh lebih rendah, bahkan kurang dari 50.000 data. Ketidakseimbangan data yang ekstrem ini, sebagaimana ditunjukkan dalam visualisasi distribusi label tersebut, dapat menyebabkan model

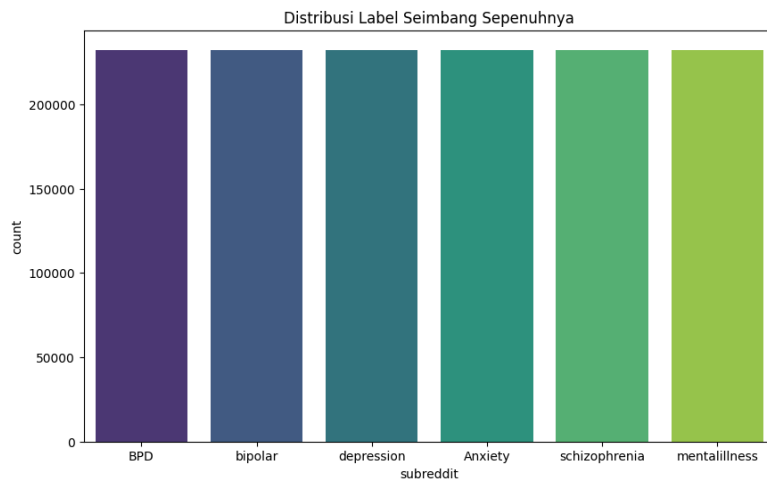
cenderung memihak pada kelas mayoritas saat proses pelatihan jika tidak ditangani.



## 2. Proses Transformasi Data

Setelah tahap eksplorasi selesai, langkah selanjutnya adalah melakukan transformasi data melalui serangkaian proses pembersihan teks (*cleaning*) secara sistematis. Peneliti menangani nilai *null* pada kolom *selftext* dengan menggantinya menjadi string kosong dan menghapus karakter non-alfanumerik serta URL yang tidak relevan. Normalisasi teks juga dilakukan dengan mengubah seluruh huruf menjadi huruf kecil (*lowercase*) untuk menjaga konsistensi fitur. Sebagai tahap akhir pembersihan, peneliti menggabungkan kolom *title* dan *selftext* menjadi satu fitur teks tunggal untuk memperkaya informasi yang akan dipelajari oleh model.

Proses transformasi data kemudian dilanjutkan dengan penerapan teknik *Easy Data Augmentation* (EDA) untuk memperkaya dataset, terutama pada kategori yang memiliki jumlah sampel sedikit. Peneliti menggunakan empat operasi utama EDA, yaitu penggantian sinonim (*synonym replacement*), penyisipan acak (*random insertion*), pertukaran acak (*random swap*), dan penghapusan acak (*random deletion*). Teknik ini diterapkan dengan aturan ketat agar perubahan yang dihasilkan tidak merusak makna asli dari ungkapan emosional yang ada dalam teks. Dampak dari proses augmentasi ini dapat diamati pada Gambar 4.2, yang menampilkan distribusi label seimbang sepenuhnya, di mana setiap kelas kini memiliki jumlah data latih yang sama besar.



Setelah data teraugmentasi, teks kemudian diubah menjadi vektor numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) agar dapat diproses oleh algoritma klasifikasi.

### 3. Hasil Modeling

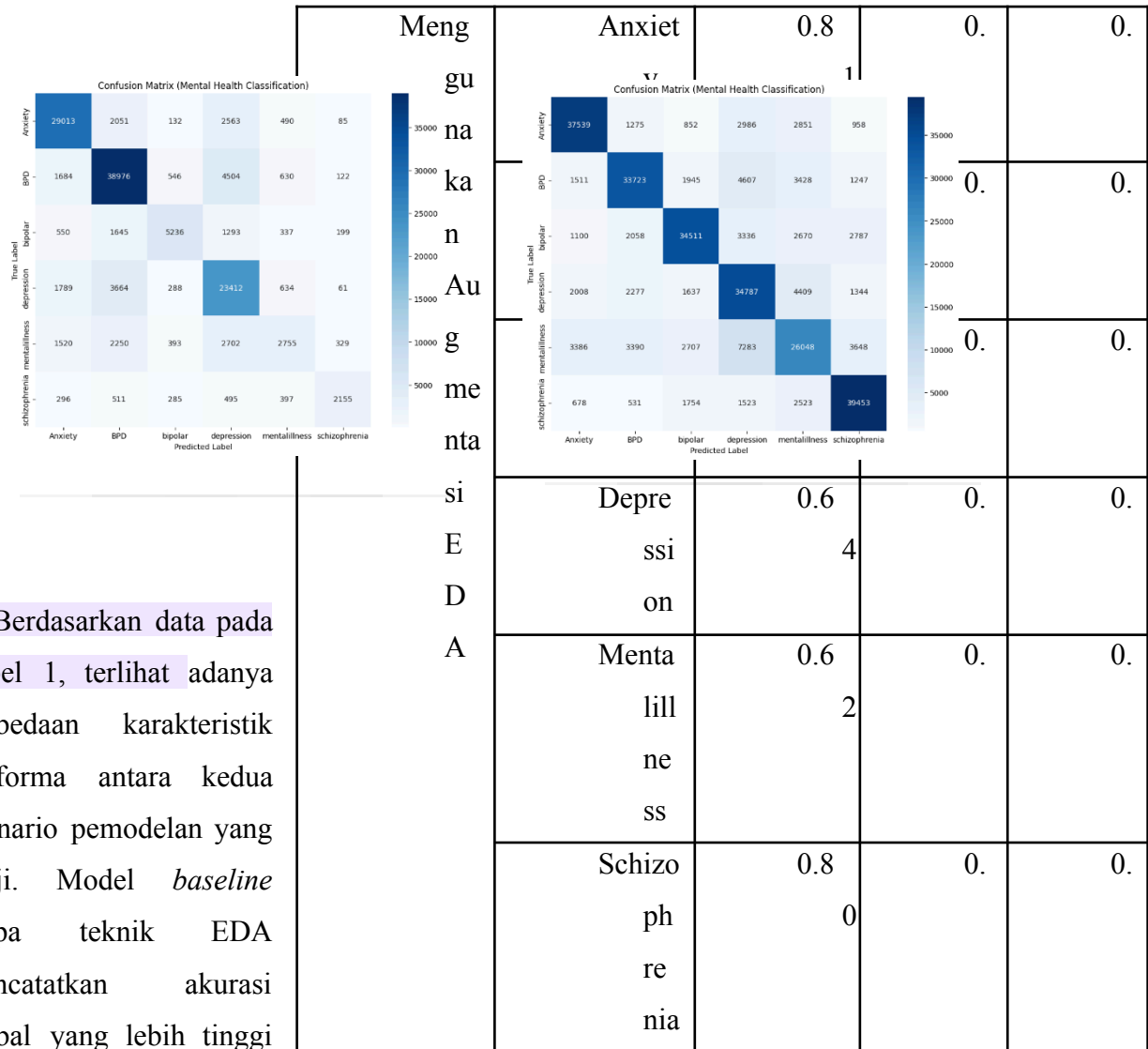
Pada tahap ini, peneliti membangun dua model klasifikasi menggunakan algoritma *Logistic Regression* sebagai basis perbandingan utama. Model pertama dilatih menggunakan dataset asli tanpa melalui proses augmentasi untuk melihat performa awal dari kombinasi TF-IDF dan *Logistic Regression*. Sementara itu, model kedua dibangun menggunakan dataset yang telah diperkaya dengan teknik EDA untuk menguji hipotesis mengenai dampak penambahan variasi teks. Kedua model tersebut kemudian dievaluasi menggunakan data uji yang telah dipisahkan sebelumnya untuk memastikan objektivitas hasil penilaian.

Hasil pengujian menunjukkan perbedaan performa antara kedua skenario pemodelan yang dilakukan peneliti. Evaluasi secara rinci divisualisasikan menggunakan *Confusion Matrix* yang dapat dilihat pada Gambar 4.3 (untuk model tanpa EDA) dan Gambar 4.4 (untuk model dengan augmentasi EDA). Pada matrik tersebut, perbedaan jumlah prediksi benar (diagonal utama) dan kesalahan prediksi (*False Positives/False Negatives*) antara kedua model dapat diamati secara

langsung. Angka-angka hasil pengujian ini kemudian dikompilasi secara komprehensif ke dalam Tabel 4.1, yang memaparkan nilai ukur *precision*, *recall*, *F1-score*, dan *accuracy* secara keseluruhan.

Tabel 4.1 Perbandingan model

|                      | Class          | Precision | R  | F  |
|----------------------|----------------|-----------|----|----|
| Tanpa Augmentasi EDA | Anxiety        | 0.83      | 0. | 0. |
|                      | BPD            | 0.79      | 0. | 0. |
|                      | Bipolar        | 0.76      | 0. | 0. |
|                      | Depression     | 0.67      | 0. | 0. |
|                      | Mental illness | 0.53      | 0. | 0. |
|                      | Schizophrenia  | 0.73      | 0. | 0. |



Berdasarkan data pada Tabel 1, terlihat adanya perbedaan karakteristik performa antara kedua skenario pemodelan yang diuji. Model *baseline* tanpa teknik EDA mencatatkan akurasi global yang lebih tinggi

sebesar 0.76 (76%) dibandingkan model dengan EDA yang mencapai 0.74 (74%). Penurunan akurasi ini merupakan dampak dari masuknya variasi teks buatan yang bersifat *noise* pada data latih, sehingga sedikit memengaruhi stabilitas prediksi pada kelas mayoritas yang sudah dominan.

Namun, jika ditinjau dari tujuan utama penelitian sebagai instrumen penyaringan (*screening*), model yang didukung teknik EDA memberikan hasil yang jauh lebih optimal secara praktis. Hal ini dibuktikan dengan lonjakan nilai *Recall* pada kelas minoritas *Mentalillness* yang meningkat signifikan dari 0.28 menjadi 0.56, serta kenaikan *F1-Score* dari 0.36 menjadi 0.59. Dalam konteks medis dan deteksi dini, kemampuan model untuk meminimalisir kesalahan deteksi pada kelas yang jarang muncul (*False Negative*) jauh lebih diprioritaskan daripada sekadar mengejar akurasi keseluruhan. Temuan ini mengonfirmasi bahwa

penerapan augmentasi EDA berhasil memperkaya pola linguistik model sehingga menjadi lebih peka dalam mendeteksi indikasi risiko gangguan kesehatan mental.

## B. Pembahasan

### 1. Analisis Model Optimal

Berdasarkan hasil pengujian yang telah dilakukan, model klasifikasi tanpa menggunakan teknik augmentasi EDA menunjukkan tingkat akurasi keseluruhan yang sedikit lebih tinggi, yaitu sebesar 0.76 (76%), dibandingkan dengan model yang menggunakan EDA yang mencapai 0.74 (74%). Penurunan akurasi global pada model dengan EDA ini kemungkinan besar disebabkan oleh efek samping penambahan variasi buatan yang justru memberikan sedikit *noise* pada pola teks di kelas mayoritas. Meskipun terjadi penurunan akurasi secara keseluruhan, model dengan penerapan teknik EDA terbukti memberikan dampak positif yang sangat signifikan terhadap kemampuan deteksi pada kelas minoritas, khususnya pada kategori *Mentalillness*. Peningkatan metrik evaluasi secara spesifik pada kelas *Mentalillness* ini menegaskan bahwa teknik augmentasi EDA berhasil mengatasi masalah keterbatasan jumlah sampel dan membantu model mengenali pola ekspresi yang lebih beragam, walaupun harus mengorbankan sedikit stabilitas performa pada kelas yang sudah dominan.

### 2. Evaluasi Terhadap Tujuan Bisnis

Evaluasi pemodelan ini dinilai berdasarkan tujuan utama bisnis, yaitu menciptakan instrumen penyaringan awal (*screening*) yang adil dan peka terhadap berbagai kategori gangguan mental. Secara praktis, penerapan teknik EDA menjadikan model jauh lebih relevan untuk kebutuhan *screening* karena berhasil menyeimbangkan sensitivitas deteksi pada kelas minoritas. Hal ini dibuktikan secara konkret pada kelas *Mentalillness*, di mana kemampuan model dalam mengenali indikasi yang benar (*Recall*) melonjak drastis dari 0.28 menjadi 0.56. Meskipun akurasi global mengalami sedikit penurunan, peningkatan kepekaan model pada kasus-kasus yang jarang muncul ini jauh lebih bernilai secara bisnis untuk meminimalisir risiko terlewatnya deteksi

pasien yang membutuhkan bantuan.

### 3. Kesesuaian Hasil dengan Ekspektasi pada Tahap Business Understanding

Hasil pemodelan yang diperoleh secara keseluruhan telah selaras dengan skenario dan ekspektasi awal yang dirumuskan pada tahap *Business Understanding*. Pada fase perencanaan tersebut, peneliti telah memproyeksikan bahwa tujuan utama teknik EDA adalah untuk menangani ketidakseimbangan kelas (*class imbalance*) dan memperkaya variasi linguistik, bukan semata-mata menaikkan akurasi global secara instan. Penurunan tipis pada akurasi keseluruhan dari 0.76 menjadi 0.74 merupakan efek samping yang sudah diantisipasi akibat masuknya *noise* buatan pada teks kelas mayoritas. Sebaliknya, lonjakan metrik *F1-Score* pada kelas *Mentalillness* dari 0.36 menjadi 0.59 menjadi bukti empiris yang mengonfirmasi bahwa ekspektasi peneliti terhadap kemampuan EDA pada kelas minoritas telah tercapai dengan baik.

### 4. Validitas dan Generalisasi Hasil

Validitas hasil dalam penelitian ini tidak hanya diukur melalui performa metrik evaluasi komputasional di atas kertas, melainkan juga dibuktikan secara empiris melalui proses validasi pakar (*expert judgment*) dengan melibatkan seorang psikolog klinis profesional. Langkah ini krusial untuk memastikan bahwa pola bahasa (*linguistic features*) dan pembobotan kata yang dipelajari oleh model NLP memiliki relevansi yang valid dengan manifestasi keluhan pasien gangguan kesehatan mental secara riil di lapangan.

Proses validasi dilakukan menggunakan instrumen kuesioner terstruktur yang terdiri dari 18 butir pertanyaan uji. Kuesioner ini menyajikan sampel teks keluhan berbahasa Indonesia yang diambil secara acak dari data uji, di mana masing-masing dari 6 kategori gangguan (*Anxiety*, *BPD*, *Bipolar*, *Depression*, *Schizophrenia*, dan *Mentalillness*) diwakili oleh 3 sampel teks. Pakar diminta untuk menilai ketepatan label prediksi sistem menggunakan skala Likert 4 tingkat (Sangat Tidak Setuju, Kurang Setuju, Setuju, Sangat Setuju). Untuk analisis kuantitatif, pilihan jawaban "Setuju" dan "Sangat Setuju" dikelompokkan sebagai prediksi yang 'Sesuai/Valid'. Rekapitulasi hasil penilaian pakar disajikan pada Tabel 4.2 berikut:

Tabel 4. 2 *Validitas*

35

| Kategori Kondisi Mental | Sangat Setuju (SS) | Setuju (S) | Kurang Setuju (KS) | Tidak Setuju (TS) | Jumlah Teks Uji | Sesuai (SS + S) | Persentase Validitas |
|-------------------------|--------------------|------------|--------------------|-------------------|-----------------|-----------------|----------------------|
| <i>Anxiety</i>          | 3                  | 0          | 0                  | 0                 | 3               | 3               | 100%                 |
| <i>BPD</i>              | 2                  | 1          | 0                  | 0                 | 3               | 3               | 100%                 |
| <i>Bipolar</i>          | 0                  | 2          | 1                  | 0                 | 3               | 2               | 66,7%                |
| <i>Depression</i>       | 3                  | 0          | 0                  | 0                 | 3               | 3               | 100%                 |
| <i>Schizophrenia</i>    | 3                  | 0          | 0                  | 0                 | 3               | 3               | 100%                 |
| <i>Mentalillness</i>    | 0                  | 2          | 1                  | 0                 | 3               | 2               | 66,7%                |
| Total                   | 11                 | 5          | 2                  | 0                 | 18              | 16              | 88,9%                |

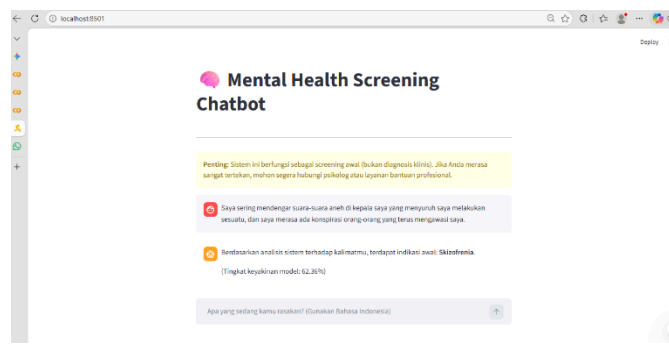
10

Berdasarkan Tabel 4.2, hasil pengujian menunjukkan rata-rata tingkat kesesuaian yang tinggi, yaitu sebesar 88,9% . Hasil ini membuktikan bahwa kombinasi metode ekstraksi fitur TF-IDF dan algoritma *Logistic Regression* yang didukung teknik augmentasi EDA berhasil menangkap ekspresi psikologis dengan akurat . Adapun ketidaksesuaian pada kelas *Bipolar* dan *Mentalillness* (masing-masing 66,7%) disebabkan oleh adanya tumpang tindih penggunaan kata (*overlapping keywords*) yang bersifat ambigu di media sosial, seperti kata "hampa" atau "stres berat" yang secara klinis bisa muncul pada beberapa kondisi gangguan mental sekaligus..

## 5. Rekomendasi Penerapan

Sebagai tindak lanjut yang merepresentasikan tahap *Deployment* pada kerangka kerja CRISP-DM, model terbaik dari hasil penelitian ini sangat direkomendasikan untuk diimplementasikan ke dalam sistem berbasis web seperti pada gambar 4.5.

dapat



Sistem tersebut nantinya

16

difungsikan sebagai instrumen pendeteksian dini mandiri, di mana pengguna dapat memasukkan teks keluhan mereka dan sistem akan menganalisis kecenderungan risiko kondisi mentalnya berdasarkan pola bahasa. Namun, peneliti merekomendasikan agar setiap penerapan sistem wajib disertai dengan (*disclaimer*) yang tegas bahwa prediksi dari model komputasional ini sebatas indikasi awal. Untuk penerapan lapangan di masa mendatang, sistem ini idealnya diintegrasikan sebagai alat bantu tambahan bagi para profesional kesehatan mental agar hasil prediksinya dapat divalidasi langsung oleh ahlinya.

## BAB V

### PENUTUP

Bab ini merupakan bagian akhir dari naskah skripsi yang memuat kesimpulan menyeluruh dari seluruh rangkaian eksperimen dan analisis hasil yang telah dilakukan untuk menjawab rumusan masalah penelitian. Selain itu, bab ini juga menyajikan saran-saran akademis maupun praktis yang direkomendasikan peneliti sebagai acuan untuk pengembangan sistem dan penelitian lanjutan di masa mendatang.

#### A. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan mengenai deteksi indikasi gangguan kesehatan mental berbasis teks dengan teknik augmentasi EDA yang mengacu pada rumusan masalah, maka dapat ditarik kesimpulan sebagai berikut:

1. Proses pengolahan dataset publik dari platform Kaggle yang bersumber dari Reddit telah berhasil dilakukan melalui serangkaian tahapan pembersihan dan praproses teks secara sistematis. Tahapan ini mencakup penanganan nilai kosong (*null*), penghapusan entri yang tidak valid seperti teks yang telah dihapus (*removed* atau *deleted*), serta pembersihan dari karakter non-alfanumerik dan URL. Selain itu, proses normalisasi dengan mengubah teks menjadi huruf kecil (*lowercase*) dan penggabungan atribut *title* beserta *selftext* dilakukan untuk memaksimalkan fitur ekstraksi. Rangkaian praproses ini terbukti krusial dalam menghasilkan dataset yang bersih, terstruktur, dan siap untuk ditransformasikan menjadi representasi fitur numerik TF-IDF untuk kebutuhan klasifikasi model.
2. Penerapan augmentasi teks menggunakan teknik *Easy Data Augmentation* (EDA) terbukti sangat efektif dalam menambah variasi ekspresi teks tanpa menuntut peneliti untuk mengumpulkan data dari sumber baru. Teknik ini diimplementasikan secara khusus pada data latih untuk memperkaya ragam linguistik, terutama pada kelas-kelas yang memiliki jumlah sampel minoritas. Melalui empat operasi utama

yang meliputi penggantian sinonim, penyisipan acak, pertukaran acak, dan penghapusan acak, dataset latih berhasil dibentuk menjadi lebih seimbang secara proporsional antar labelnya. Langkah strategis ini secara langsung telah membantu mengatasi masalah ketidakseimbangan kelas (*class imbalance*) dan mencegah model dari kecenderungan memihak pada kelas mayoritas selama proses pelatihan.

3. Perbandingan performa model *baseline Logistic Regression* menunjukkan dinamika hasil yang sangat menarik setelah dilakukan penerapan teknik augmentasi EDA berdasarkan evaluasi metrik standar. Model tanpa EDA memang mencatat akurasi global yang sedikit lebih tinggi yaitu 0.76 (76%) dibandingkan model dengan EDA yang berada pada angka 0.74 (74%). Namun demikian, penerapan EDA berhasil memberikan lonjakan performa yang sangat signifikan pada kelas minoritas, di mana nilai *Recall* untuk kategori *Mentalillness* naik drastis dari 0.28 menjadi 0.56, serta *F1-Score* yang meningkat dari 0.36 menjadi 0.59. Fakta ini menyimpulkan bahwa penggunaan model yang didukung EDA jauh lebih direkomendasikan dan relevan untuk tujuan penyaringan (*screening*) kesehatan mental, karena kemampuan mendeteksi kelas minoritas secara adil jauh lebih krusial daripada sekadar mengejar tingkat akurasi keseluruhan.

## B. Saran

Berdasarkan hasil analisis, kesimpulan, serta keterbatasan yang ditemukan selama proses penelitian berlangsung, peneliti memberikan beberapa saran untuk pengembangan selanjutnya:

1. Penelitian lanjutan sangat disarankan untuk mengaplikasikan alur kerja metode CRISP-DM ini secara langsung pada dataset asli berbahasa Indonesia yang dikumpulkan dari platform media sosial lokal. Penggunaan dataset berbahasa Indonesia secara langsung akan jauh lebih akurat dalam mencerminkan ragam bahasa informal, *slang*, serta konteks sosio-kultural masyarakat Indonesia yang sangat memengaruhi cara individu mengekspresikan kondisi mentalnya. Selain itu, peneliti

selanjutnya dapat mencoba mengeksplorasi penggunaan model algoritma *Deep Learning* seperti *Bidirectional Long Short-Term Memory* (Bi-LSTM) atau *Transformer* (misalnya IndoBERT) sebagai perbandingan. Eksplorasi pada algoritma yang lebih kompleks tersebut diharapkan mampu memahami konteks dan penalaran kalimat yang lebih dalam sehingga dapat meningkatkan performa akurasi klasifikasi secara lebih optimal.

2. Dalam aspek pengembangan dan implementasi sistem, penelitian berikutnya sangat dianjurkan untuk terus mempertahankan dan meningkatkan keterlibatan tenaga profesional seperti psikolog secara lebih komprehensif. Validasi dari pakar psikologi sangat dibutuhkan tidak hanya pada akhir evaluasi prediksi (*expert judgment*), tetapi juga disarankan ikut serta sejak tahap pelabelan data agar validitas sistem secara klinis semakin kuat. Selanjutnya, antarmuka aplikasi web (*deployment*) yang telah dirancang menggunakan Streamlit dapat dikembangkan menjadi sebuah purwarupa yang terintegrasi dengan layanan rujukan ke fasilitas kesehatan nyata. Terlepas dari inovasi yang akan dibangun, setiap bentuk implementasi sistem harus tetap memberikan peringatan atau penafian (*disclaimer*) yang tegas kepada pengguna bahwa prediksi mesin ini mutlak hanya sebagai penyaringan awal dan bukan pengganti diagnosis medis resmi.

## 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/tac1\\_a\\_00726](https://doi.org/10.1162/tac1_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., & Najibullo, 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>.

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>.

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>

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>