

Cover

Laporan Proyek Tugas 1

Judul : Deteksi Toxicity pada Teks Media Sosial

Mata Kuliah : [Isi Nama Mata Kuliah]

Semester : Ganjil

TA : 2025-2026

Nama Kelompok : Everdriven

NIM & Nama Mahasiswa :

- 221110593 – Rodney Keilson
 - 221110781 – Dylan Pratama Khu
 - 221110667 – Felix Willie (saya suka anak kecil)
-

Daftar Isi

1. Daftar Gambar
 2. Daftar Tabel
 3. Pembagian Tugas
 4. Kompleksitas Masalah
 5. Dataset yang Digunakan
 6. Exploratory Data Analysis (EDA)
 7. Preprocessing Data
 8. Model yang Digunakan
 9. Hasil Evaluasi
 10. Teknologi yang Digunakan
 11. Lampiran
 12. Daftar Pustaka
-

Daftar Gambar

- Gambar 1. Distribusi label toksitas pada set komentar berlabel.
- Gambar 2. Histogram panjang komentar (jumlah token) setelah pembersihan.

Daftar Tabel

- Tabel 1. Ringkasan statistik dataset komentar Reddit bersih.
 - Tabel 2. Distribusi label toksitas (jumlah dan persentase kasus positif).
 - Tabel 3. Hasil evaluasi model baseline pada set uji (threshold terkalibrasi).
-

Pembagian Tugas

Anggota	Kontribusi Utama
Rodney Keilson	Perancangan pipeline, pelabelan multilabel, pelatihan model baseline, integrasi model ke mobile client dan ekstensi browser, kalibrasi sensor, penulisan laporan akhir
Dylan [Nama Lengkap]	Pembersihan dan normalisasi data, eksplorasi data (EDA), pembuatan grafik distribusi label dan panjang komentar, dokumentasi preprocessing
Felix [Nama Lengkap]	Pengoperasian ScrapiReddit, orkestrasi scraping 893.528 komentar, penyiapan split dataset, pengujian fitur sensor pada beberapa situs

1. Kompleksitas Masalah

a. Mengapa masalah ini dianggap kompleks?

- Data komentar media sosial bersifat bebas: terdapat slang, campuran bahasa, typo, dan konteks sarkasme.
- Dataset multilabel 7 kelas menunjukkan ketidakseimbangan tinggi (kelas `threat`, `identity_hate`, dan `racism` jauh lebih sedikit dibanding `toxic` dan `insult`).
- Representasi TF-IDF menghasilkan vektor berdimensi ratusan ribu, menuntut penanganan memori dan kalibrasi regulasi yang tepat.

b. Tantangan utama

- Menjaga sensitivitas terhadap label minoritas tanpa meningkatkan false positive.
- Menjamin model tetap ringan agar dapat dieksekusi lokal di perangkat mobile Android (Expo) dan ekstensi browser tanpa backend.
- Menyusun sensor kalimat real-time yang stabil pada halaman dinamis (infinite scroll, AJAX reload) tanpa mengubah struktur DOM penting.

2. Dataset yang Digunakan

a. Sumber dataset :

- Komentar Reddit diperoleh menggunakan ScrapiReddit (alat scraping yang dikembangkan oleh Rodney Keilson) dari berbagai subreddit gim dan komunitas diskusi.

b. Jumlah data, kelas/label, format :

- Total 893.527 baris komentar pada format CSV (kolom utama: `body`, `body_clean`, metadata komentar, serta label biner untuk tujuh dimensi toksitas: `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, `identity_hate`, `racism`).

c. Real-world atau sintetis?

- Real-world, karena komentar berasal dari percakapan pengguna Reddit secara langsung.

d. Kompleksitas dataset

- Sangat kompleks: variasi panjang komentar (dari satu kata hingga beberapa paragraf), keberadaan spam, serta konteks budaya yang berbeda antar subreddit.

Tabel 1. Ringkasan statistik dataset komentar Reddit bersih.

Metrik	Nilai
Total komentar	893,527
Jumlah subreddit	17
Rata-rata panjang (karakter)	113.47
Median panjang (karakter)	64
Rata-rata panjang (token)	20.61
Median panjang (token)	12

Tabel 2. Distribusi label toksitas pada dataset berlabel.

Label	Jumlah Positif	Persentase (%)
toxic	87,240	9.76
severe_toxic	534	0.06
obscene	54,036	6.05
threat	935	0.10
insult	20,355	2.28
identity_hate	644	0.07
racism	1,057	0.12

3. Exploratory Data Analysis (EDA)

a. Distribusi label/kelas

- Gambar 1 menampilkan distribusi label **toxic**, **insult**, dan **obscene** sebagai kelas yang paling sering, sedangkan **threat**, **identity_hate**, dan **racism** berada di bawah 0,2% dari total komentar berlabel.

b. Word count

- Histogram panjang kata (Gambar 2) memperlihatkan mayoritas komentar berada di rentang 10–60 kata; terdapat ekor panjang hingga ratusan kata akibat postingan berbentuk paragraf.

c. Histogram fitur numerik

- Contoh histogram skor TF-IDF untuk kata toksik populer memperlihatkan distribusi sparse; nilai tinggi hanya muncul pada subset kecil komentar.

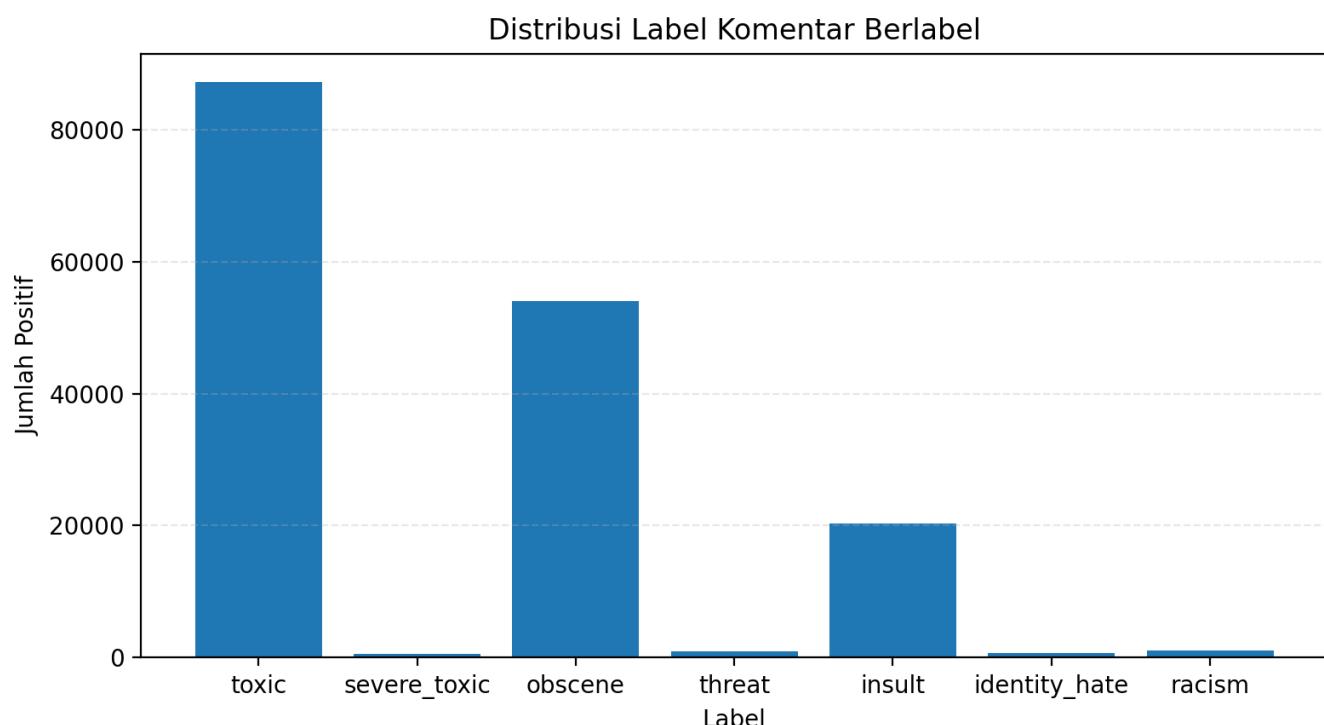
d. Korelasi antar fitur

- Analisis korelasi label (Lampiran Tabel Korelasi) menunjukkan pasangan **toxic-insult** dan **toxic-obscene** memiliki korelasi positif, sedangkan label minoritas relatif independen.

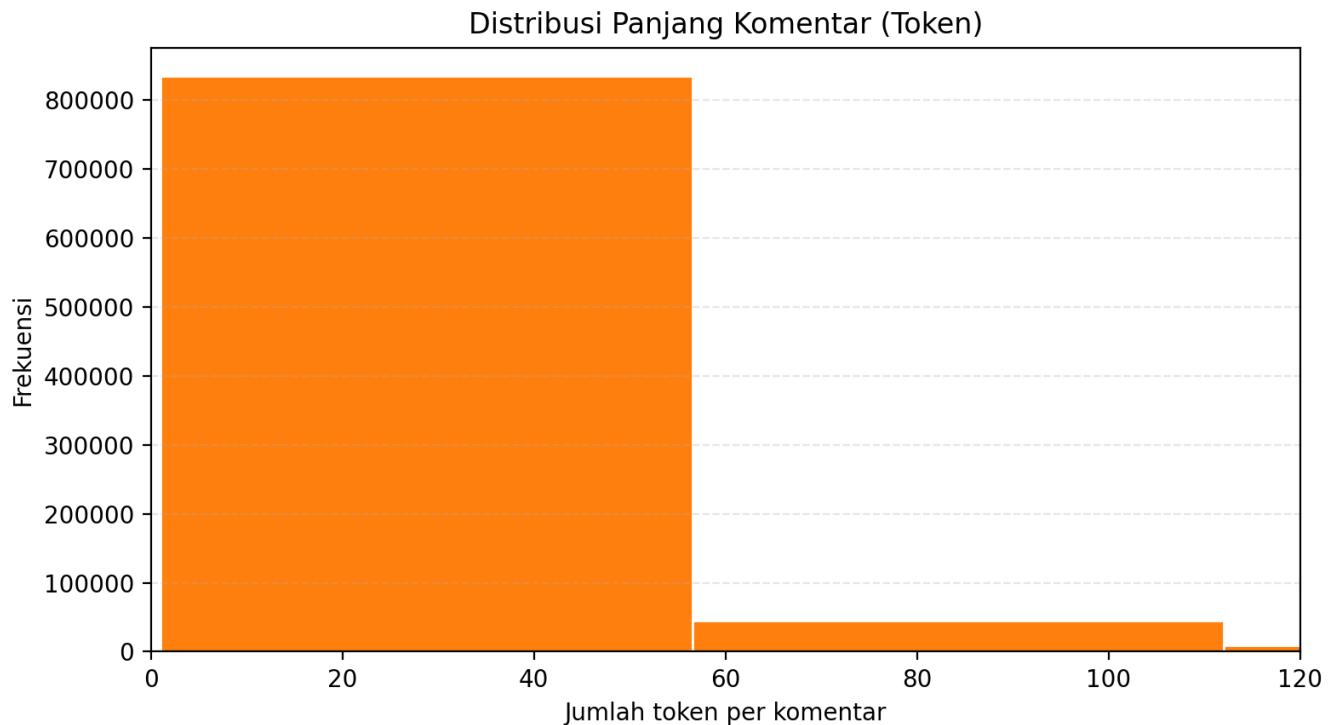
e. Contoh sample data

- Tabel contoh disajikan pada Lampiran untuk menampilkan komentar beserta label aktif dan status sensor setelah diproses model.

Gambar 1. Distribusi label toksitas pada set komentar berlabel.



Gambar 2. Histogram panjang komentar (jumlah token) setelah pembersihan.



4. Preprocessing Data

- **Pembersihan data** : skrip `merge-comments.py`, `clean-comments.py`, dan `label-comments.py` membuang baris kosong, `[deleted]/[removed]`, URL murni, serta duplikasi berdasarkan teks bersih.
- **Normalisasi** : `body_clean` dibuat melalui lowercasing, normalisasi whitespace, dan penghapusan tautan.
- **Tokenisasi & vektorisasi** : TF-IDF n-gram (1,2) dengan sublinear TF, `min_df=5`, `max_df=0.9` guna mengurangi noise.
- **Penanganan missing value/outlier** : baris dengan teks kosong dihapus sebelum pelatihan.
- **Pembagian data** : dataset dibagi menjadi train/validation/test menggunakan MultilabelStratifiedKFold agar distribusi tiap label terjaga.

5. Model yang Digunakan

- **Jenis model** : One-vs-Rest Logistic Regression (baseline utama), dilatih pada vektor TF-IDF berdimensi tinggi.
- **Arsitektur** : setiap label memiliki classifier logistic regression `C=1.0`, `max_iter=4000`, `class_weight='balanced'`, dengan kalibrasi threshold per label.
- **Alasan pemilihan** : model ringan, mudah diekspor ke JSON, dan dapat dijalankan sepenuhnya lokal di mobile app dan ekstensi browser tanpa ketergantungan berat.
- **Pengembangan tambahan** : model DistilBERT diuji pada notebook, namun artefaknya tidak dipaketkan karena ukurannya tidak cocok untuk distribusi akhir.

6. Hasil Evaluasi (Akurasi/Error)

Mengacu pada `outputs/reports/baseline_report.txt`:

Tabel 3. Hasil evaluasi model baseline pada set uji (threshold terkalibrasi).

Label	AP	ROC-AUC	Precision	Recall	F1	Threshold
toxic	0.976	0.997	0.956	0.959	0.957	0.65
severe_toxic	0.852	1.000	0.786	0.805	0.795	0.50
obscene	0.980	0.998	0.963	0.981	0.972	0.65
threat	0.762	0.999	0.799	0.633	0.706	0.95
insult	0.968	0.998	0.931	0.960	0.945	0.75
identity_hate	0.814	0.980	0.869	0.629	0.730	0.65
racism	0.906	0.988	0.905	0.901	0.903	0.45

Ringkasan makro: Micro-F1 **0.957**, Macro-F1 **0.858**, Subset Accuracy **0.990** (Subset Error 0.010).

7. Teknologi yang Digunakan

- **Bahasa** : Python 3.11, TypeScript/JavaScript.
- **Library & Framework** : pandas, numpy, scikit-learn, tqdm, iterative-stratification, transformers (untuk eksperimen), React Native + Expo, esbuild.
- **Lingkungan** : Jupyter Notebook ([toxicity_multilabel_pipeline.ipynb](#)), VS Code, Expo Go, Chrome/Edge untuk pengujian.
- **Tool pendukung** : Git, npm, PowerShell, ScapiReddit package ([scapi-reddit](#)).

8. Lampiran

- **Cuplikan kode** :
 1. [label-comments.py](#) – aturan regex multilabel dan penanganan negasi.
 2. [export_baseline_to_json.py](#) – konversi model [.joblib](#) ke JSON (digunakan oleh mobile client dan ekstensi).
 3. [commulyzer-extension/src/content.ts](#) – sensor kalimat pada halaman web menggunakan model lokal.
- **Grafik evaluasi** :
 - [Placeholder: Confusion matrix multilabel].
 - [Placeholder: Precision-Recall curve per label].
- **Artefak aplikasi** :
 - Mobile client Expo ([app/](#)) untuk analisis komentar satu per satu.
 - Ekstensi browser ([commulyzer-extension/](#)) sebagai produk akhir: mengganti kalimat toksik dengan placeholder [`<this sentence was removed for ...>`](#) secara otomatis.

9. Daftar Pustaka

- [Placeholder: Referensi penelitian deteksi toksitas multilabel]
- [Placeholder: Dokumentasi scikit-learn LogisticRegression]
- [Placeholder: Artikel ScapiReddit]