



Hate Speech and Offensive Language Classification Using Hybrid RoBERTa and XGBoost with Hyperparameter Optimization

Klasifikasi Hate Speech dan Offensive Language Menggunakan Hybrid RoBERTa dan XGBoost dengan Optimasi Hyperparameter

Najwa Jauhari^{1*}, Surya Agustian², Fadhilah Syafria³, Muhammad Affandes⁴

¹²³⁴Program Studi Teknik Informatika, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

E-Mail: ¹12250124838@students.uin-suska.ac.id, ²surya.agustian@uin-suska.ac.id,
³fadhilah.syafria@uin-suska.ac.id, ⁴affandes@uin-suska.ac.id

Received May 4th 2026; Revised Jun 05th 2026; Accepted Jun 21th 2026; Available Online Jun 25th 2026

Corresponding Author: Najwa Jauhari

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Hate speech and offensive language on social media require accurate automatic detection. This study proposes a hybrid approach combining a Twitter-based Robustly Optimized BERT Approach (RoBERTa) model (cardiffnlp/twitter-roberta-base-offensive) as a feature extractor with XGBoost as a classifier on the English Hate Speech and Offensive Content Identification (HASOC) 2021 dataset. The novelty lies in integrating Optuna-based hyperparameter optimization, multi-seed ensemble, division calibration, and oversampling within the feature embedding space, which has never been applied simultaneously in a single pipeline for HASOC 2021. Two tasks are addressed: Task 1A is a binary classification task that distinguishes Hate and Offensive (HOF) content from Non-Hate and Offensive (NOT) content. Task 1B is a multi-class classification task that categorizes tweets into four classes: hate speech (HATE), Offensive Language (OFFN), Profane Language (PRFN), and Safe Content (NONE). Evaluation on the test set yields Macro F1-Scores of 0.8083 for Task 1A and 0.6541 for Task 1B. Comparison with the official HASOC 2021 leaderboard shows that these scores are comparable to the 5th-ranked (Task 1A) and 3rd-ranked (Task 1B) teams using pure RoBERTa-based approaches. The main contribution is the hybrid RoBERTa-XGBoost architecture and optimization strategies, achieving competitive performance compared to Transformer-based models.

Keywords: HASOC, Hate Speech, Hyperparameter Optimization, Offensive Language, RoBERTa, XGBoost

Abstrak

Ujaran kebencian dan bahasa ofensif di media sosial merupakan masalah serius yang membutuhkan deteksi otomatis yang akurat. Penelitian ini mengusulkan pendekatan hybrid yang menggabungkan model *Robustly Optimised BERT Approach* (RoBERTa) berbasis Twitter (cardiffnlp/twitter-roberta-base-offensive) sebagai ekstraktor fitur dengan algoritma XGBoost untuk klasifikasi pada dataset *Hate Speech and Offensive Content Identification* (HASOC) 2021 berbahasa Inggris. Kebaruan penelitian ini terletak pada integrasi optimasi *hyperparameter Optuna*, *Multi-Seed Ensemble*, *Division Calibration*, dan *oversampling* pada ruang fitur *embedding* yang belum pernah diterapkan secara bersamaan dalam satu pipeline pada HASOC 2021. Dua tugas diselesaikan : Task 1A adalah klasifikasi biner untuk membedakan konten *Hate and Offensive* (HOF) dari yang tidak (NOT). Task 1B adalah klasifikasi multi-kelas yang membagi tweet menjadi empat kategori: (ujaran kebencian terhadap kelompok) HATE, bahasa ofensif terhadap individu (OFFN), kata kasar tanpa target spesifik (PRFN), dan konten aman (NONE). Ketidakseimbangan kelas ditangani dengan *oversampling* pada ruang fitur, dan optimasi *hyperparameter* dilakukan untuk meningkatkan performa. Hasil evaluasi pada data uji menunjukkan Macro F1-Score sebesar 0,8083 untuk Task 1A dan 0,6541 untuk Task 1B. Perbandingan dengan papan peringkat HASOC 2021 menunjukkan bahwa skor tersebut sebanding dengan tim peringkat 5 (Task 1A) dan peringkat 3 (Task 1B) yang berbasis BERT murni. Kontribusi utama adalah arsitektur hybrid RoBERTa-XGBoost serta strategi optimasi yang kompetitif dibandingkan model Transformer.

Kata Kunci: HASOC, Hate Speech, Optimasi Hyperparameter, RoBERTa, Ujaran Kebencian, XGBoost

1. PENDAHULUAN

Perkembangan media sosial, khususnya X (sebelumnya Twitter), telah membawa dampak signifikan terhadap penyebaran *hate speech* dan *offensive language* secara global. Sekretaris Jenderal PBB António Guterres menyatakan bahwa media sosial yang tidak terkendali memungkinkan *hate speech* menyebar seperti api dan memicu kekerasan [1]. Laporan transparansi X periode Januari-Juni 2024 mencatat lebih dari 66,9 juta aduan terkait konten kebencian, menjadikannya kategori pelanggaran tertinggi kedua di platform tersebut [2], [3]. Dampak yang ditimbulkan bersifat multidimensional: kajian sistematis Madriaza *et al.* [4] membuktikan bahwa paparan ujaran kebencian memperburuk kesejahteraan psikologis individu secara signifikan, ditandai dengan munculnya gejala depresi ($d = 1,105$; $p < 0,01$), penurunan kepuasan hidup, serta meningkatnya rasa takut sosial dan erosi kepercayaan antarkelompok. Survei global University of Oxford terhadap 10 negara demokrasi juga mengungkapkan bahwa 59% responden menganggap paparan terhadap konten kebencian tidak dapat dihindari di media sosial, dan 79% mendukung penghapusan hasutan kekerasan dari platform digital [5]. Di Indonesia, penyebaran konten kebencian berbasis Suku, Agama, Ras, dan Antargolongan (SARA) dapat dikenai sanksi pidana berdasarkan Pasal 28 ayat (2) jo. Pasal 45A ayat (2) UU ITE dengan ancaman penjara hingga enam tahun [6]. Besarnya skala dampak ini menegaskan urgensi pengembangan sistem deteksi otomatis yang akurat sebagai bagian dari infrastruktur moderasi konten digital.

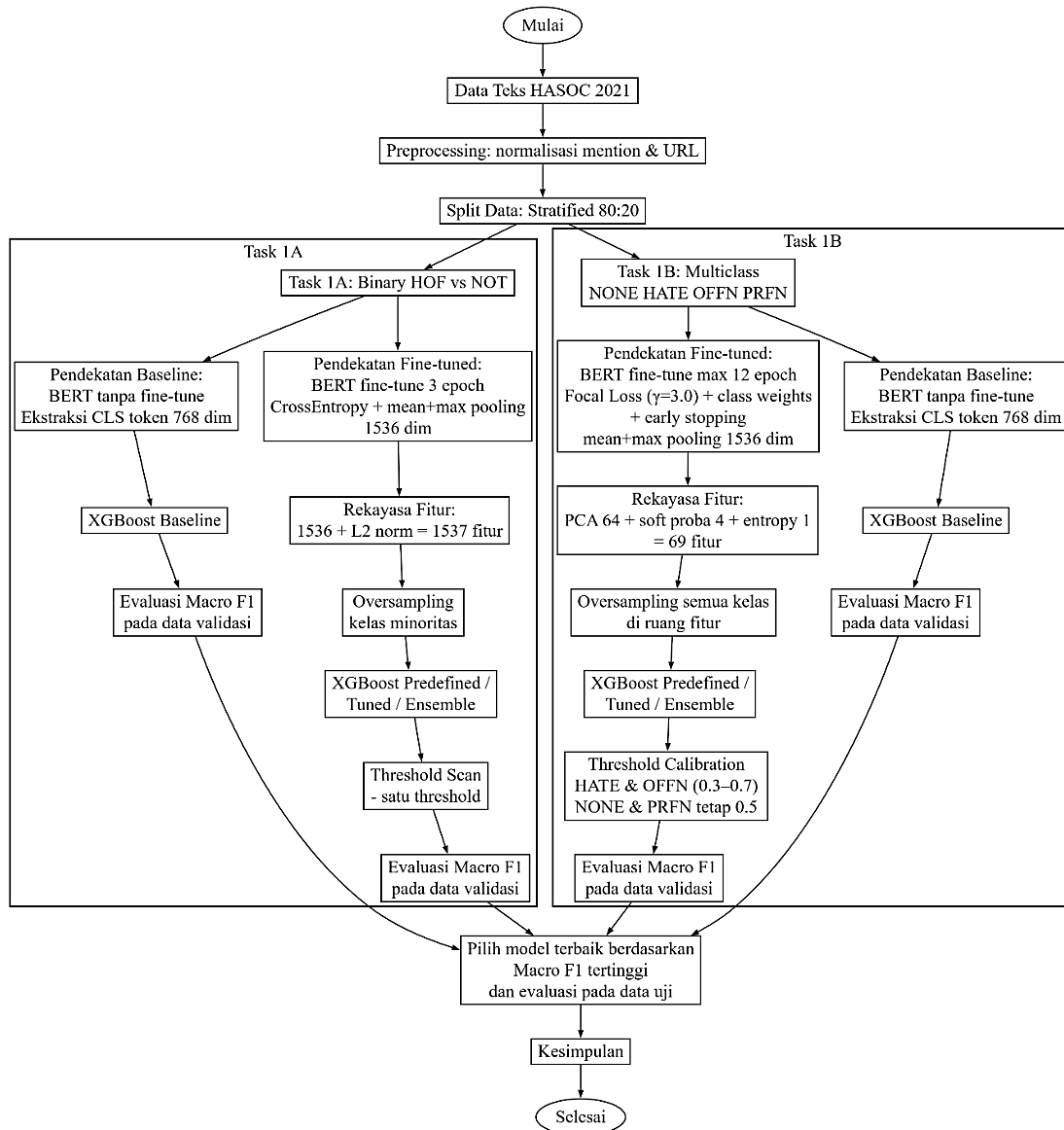
Berbagai *shared task* telah diselenggarakan untuk mendorong riset deteksi *hate speech* secara otomatis, salah satunya *Hate Speech and Offensive Content Identification* (HASOC) yang telah berjalan sejak 2019 [7], [8], [9]. Dataset HASOC 2021 untuk bahasa Inggris dikumpulkan dari Twitter dan disediakan melalui *Forum for Information Retrieval Evaluation* (FIRE) 2021, terdiri dari 3.843 tweet latih dan 1.281 tweet uji dengan anotasi manual. Dataset ini mencakup dua skema klasifikasi. Task 1A adalah klasifikasi biner untuk membedakan konten *Hate and Offensive* (HOF) dari yang tidak (NOT). Task 1B adalah klasifikasi multi-kelas yang membagi tweet menjadi empat kategori: (ujaran kebencian terhadap kelompok) HATE, bahasa ofensif terhadap individu (OFFN), kata kasar tanpa target spesifik (PRFN), dan konten aman (NONE). Distribusi data menunjukkan ketidakseimbangan kelas yang signifikan, dengan rasio HOF terhadap NOT sekitar 1,9:1 pada Task 1A, sementara kelas HATE dan OFFN menjadi minoritas pada Task 1B. Di luar persoalan distribusi kelas, teks media sosial menghadirkan tantangan linguistik yang kompleks, mencakup penggunaan bahasa informal, *slang*, singkatan, serta fenomena pragmatik seperti ironi dan sarkasme yang menyulitkan sistem dalam memahami makna sebenarnya dari sebuah ujaran. Konteks linguistik yang ambigu ini menjadikan deteksi *hate speech* pada data Twitter jauh lebih kompleks dibandingkan dengan klasifikasi teks formal pada umumnya.

Penelitian deteksi *hate speech* telah berkembang melalui berbagai pendekatan. Metode *machine learning* klasik seperti *Extreme Gradient Boosting* (XGBoost) dengan fitur *Term Frequency-Inverse Document Frequency* (TF-IDF) tergolong efisien, namun kurang mampu menangkap konteks semantik [10]. *Deep learning* berbasis *Convolutional Neural Network* (CNN) dan *Long Short-Term Memory* (LSTM) dapat menangkap pola urutan kata, tetapi masih terbatas dalam memahami ketergantungan jarak jauh dan konteks dua arah. Model Transformer seperti *Bidirectional Encoder Representations from Transformers* (BERT) dan *Robustly Optimized BERT Pretraining Approach* (RoBERTa) menawarkan representasi kontekstual yang lebih kaya [11], [12]; misalnya, Almaliki *et al.* [13] mencapai akurasi 98,6% menggunakan BERT-Mini untuk bahasa Arab. Meski demikian, Transformer murni memiliki kelemahan berupa biaya komputasi tinggi, rawan *overfitting* pada dataset kecil, serta relatif sulit di-deploy. Pendekatan *hybrid* yang menggabungkan Transformer sebagai ekstraktor fitur dengan *classifier* klasik mulai dieksplorasi sebagai alternatif yang lebih praktis. Salminen *et al.* [14] melaporkan bahwa kombinasi XGBoost dengan fitur BERT mencapai F1-Score 0,924 lintas platform, sedangkan Iftikhar *et al.* [15] menunjukkan bahwa *pipeline* RoBERTa-large + XGBoost mengungguli CNN maupun RoBERTa tunggal. Ashiq *et al.* [16] juga melaporkan performa kompetitif pada bahasa Roman Urdu menggunakan pendekatan serupa. Namun, studi-studi tersebut belum mengeksplorasi strategi optimasi yang komprehensif pada dataset kompetisi standar seperti HASOC.

Berdasarkan penelusuran terhadap 46 laporan peserta HASOC 2021 [9], tidak ditemukan satu pun pendekatan yang mengombinasikan RoBERTa sebagai *feature extractor* dengan XGBoost sebagai *classifier*; mayoritas peserta menggunakan *fine-tuned* Transformer murni atau kombinasi dengan CNN/LSTM, dan satu-satunya tim yang menggunakan XGBoost justru mengombinasikannya dengan fitur TF-IDF [10]. Selain itu, belum ada studi yang secara bersamaan menerapkan optimasi *hyperparameter* berbasis Optuna, *multi-seed ensemble*, dan *division calibration* dalam satu *pipeline* terintegrasi untuk dataset HASOC. *Gap* inilah yang menjadi landasan penelitian ini. Penelitian ini mengusulkan model *hybrid* yang memanfaatkan *cardiffnlp/twitter-roberta-base-offensive* [17] sebagai *feature extractor*, dikombinasikan dengan XGBoost yang dioptimasi menggunakan Optuna, *multi-seed ensemble*, *division calibration*, serta *oversampling* kustom pada ruang fitur. Tujuan penelitian ini adalah membangun dan mengevaluasi model klasifikasi *hate speech* dan *offensive language* pada dataset HASOC 2021 (Task 1A dan Task 1B), dengan menguji apakah strategi optimasi berlapis tersebut mampu menghasilkan performa yang kompetitif dan stabil dibandingkan *baseline* maupun peserta kompetisi HASOC 2021.

2. METODOLOGI PENELITIAN

Metode yang digunakan dalam penelitian ini adalah pendekatan hybrid yang menggabungkan *Robustly Optimized BERT Pretraining Approach* (RoBERTa) sebagai *feature extractor* dengan *Extreme Gradient Boosting* (XGBoost) sebagai *classifier*, diimplementasikan pada dataset *Hate Speech and Offensive Content Identification* (HASOC) 2021. Secara keseluruhan, alur penelitian terdiri dari beberapa tahap utama: pra-pemrosesan teks, ekstraksi fitur menggunakan RoBERTa, penanganan ketidakseimbangan kelas, serta pelatihan dan optimasi XGBoost. Seluruh tahapan dijalankan secara terpisah untuk Task 1A (biner) dan Task 1B (multi-kelas), dengan evaluasi akhir pada data uji HASOC 2021. Gambar 1 mengilustrasikan alur tersebut.



Gambar 1. Tahapan Penelitian

Secara garis besar, penelitian ini menggunakan dataset HASOC 2021 berbahasa Inggris yang terdiri dari 3.843 tweet latih dan 1.281 tweet uji. Proses diawali dengan pra-pemrosesan teks untuk menormalisasi *mention* dan URL. Data kemudian dibagi menjadi data latih dan validasi menggunakan *stratified split* 80:20. Selanjutnya, proses dijalankan secara terpisah untuk dua tugas klasifikasi.

Pada Task 1A yaitu klasifikasi biner *Hate and Offensive* (HOF) dan yang tidak (NOT), terdapat dua jalur pendekatan. Jalur pertama adalah *baseline* yang menggunakan BERT tanpa *fine-tuning* dengan ekstraksi token [CLS] (768 dimensi) dan dilanjutkan dengan XGBoost *baseline*. Jalur kedua menggunakan RoBERTa yang di-*fine-tune* dengan ekstraksi fitur *dual pooling* (mean+max) dan fitur L2-norm, menghasilkan 1537 fitur, kemudian dilatih dengan XGBoost dalam tiga konfigurasi: *predefined* (parameter default), *tuned* (optimasi Optuna), dan *ensemble* (5 seed). Prediksi akhir pada jalur kedua ditentukan melalui *threshold scan* untuk memilih ambang batas probabilitas terbaik. Model terbaik dipilih berdasarkan *Macro F1* tertinggi pada data validasi, lalu dievaluasi pada data uji.

Pada Task 1B (klasifikasi multi-kelas), RoBERTa di-*fine-tune* menggunakan *Focal Loss* ($\gamma=3.0$) dengan *class weights* dan *early stopping*. Ekstraksi fitur menggunakan *dual pooling* (mean+max) menghasilkan 1536 dimensi, kemudian direduksi dengan *Principal component analysis* (PCA) menjadi 64 komponen, ditambah *soft probability* (4 dimensi) dan *entropy* (1 dimensi) sehingga total 69 fitur. *Oversampling* dilakukan pada ruang fitur untuk semua kelas minoritas. *Extreme Gradient Boosting* (XGBoost) dilatih dalam tiga konfigurasi: *predefined*, *tuned* (Optuna), dan *ensemble* (5 seed). *Threshold calibration* diterapkan dengan *grid search* pada rentang 0,3–0,7 untuk kelas (ujaran kebencian terhadap kelompok) HATE, bahasa ofensif terhadap individu (OFFN), sementara NONE dan PRFN tetap 0,5. Model terbaik dipilih berdasarkan *Macro F1* tertinggi pada data validasi, lalu dievaluasi pada data uji.

2.1. Pengumpulan Data

Dataset yang digunakan adalah HASOC 2021 untuk bahasa Inggris, yang disediakan melalui kompetisi HASOC 2021 sebagai salah satu *shared task* dalam Forum for Information Retrieval Evaluation (FIRE). Dataset terdiri dari 3.843 *tweet* untuk pelatihan dan 1.281 *tweet* untuk data uji, dengan dua skema klasifikasi: Task 1A (biner: NOT dan HOF) serta Task 1B (multi-kelas: PRFN, HATE, OFFN, dan NONE). Tabel 1 menyajikan distribusi data pelatihan dan data uji.

Tabel 1. Data Train dan Data Testing HASOC 2021 Bahasa Inggris

Task	Kelas	Keterangan	Training	Testing
Task 1A	NOT	Non Hate-Offensive	1.342	798
	HOF	Hate and Offensive	2.501	483
Task 1B	NONE	Non Hate-Offensive	1.342	798
	HATE	Hate Speech	683	224
	OFFN	Offensive Language	622	195
	PRFN	Profane	1.196	379
Total			3.843	1.281

Berdasarkan Tabel 1, data *training* menunjukkan ketidakseimbangan kelas yang signifikan. Pada Task 1A, kelas HOF (2.501 sampel) mendominasi dibandingkan dengan NOT (1.342 sampel) dengan rasio sekitar 1,9:1. Pada Task 1B, kelas HATE (683) dan OFFN (622) merupakan kelas minoritas. Seluruh data *training* digunakan dalam proses pengembangan model, sedangkan 1.281 data *testing* yang disediakan penyelenggara digunakan sebagai data evaluasi akhir yang terpisah.

2.2. Pra-Pemrosesan Data

Pra-pemrosesan teks dilakukan secara minimal mengikuti panduan resmi dari model cardiffnlp/twitter roberta base offensive [17]. Hanya dua langkah normalisasi yang diterapkan:

1. Normalisasi mention pengguna - setiap mention yang diawali @ dan memiliki panjang lebih dari satu karakter diubah menjadi token generik @user.
2. Normalisasi URL - semua tautan yang diawali dengan http diganti dengan token http.

Pendekatan ini mempertahankan emoji, *hashtag*, dan kapitalisasi karena model telah mempelajari elemen tersebut saat *pretraining* di Twitter, sehingga konteks emosional tetap terjaga.

Tabel 2. Contoh *Tweet* per Kelas pada Dataset HASOC 2021

No.	Langkah	Sebelum	Sesudah
1	Normalisasi mention	@john_doe you are stupid	@user you are stupid
2	Normalisasi URL	Check this https://t.co/xyz	Check this http

Kedua langkah di atas diterapkan secara berurutan pada seluruh data *training*, *validasi*, dan *testing*. Proses ini menjaga keseimbangan antara menyederhanakan variasi teks yang tidak informatif (seperti nama pengguna dan URL spesifik) dan tetap mempertahankan konteks semantik yang kaya untuk model.

2.3. Split Data

Pembagian data dilakukan menggunakan *stratified split* 80:20 dengan *random state* = 42. Metode ini menjaga proporsi setiap kelas tetap sama di subset *training* maupun *validasi*, sehingga evaluasi lebih representatif terhadap data yang tidak seimbang. Dari total 3.843 data *training* awal, dihasilkan 3.074 data *training* (80%) untuk melatih model dan 769 data *validasi* (20%) untuk memantau performa selama pengembangan. Sementara itu, 1.281 data *testing* yang disediakan oleh penyelenggara HASOC 2021 digunakan sebagai data pengujian akhir yang terpisah dan tidak dilibatkan sama sekali dalam proses pelatihan maupun validasi.

2.4. Ekstraksi Fitur

2.4.1. RoBERTa sebagai Feature Extractor

Model yang digunakan adalah varian RoBERTa berbasis Twitter (cardiffnlp/twitter-roberta-base-offensive), yang merupakan pengembangan dari arsitektur BERT. Model ini telah dilatih pada sekitar 58 juta tweet dan di-*fine-tune* untuk mengidentifikasi bahasa ofensif pada *benchmark* TweetEval [18]. RoBERTa dibangun di atas arsitektur BERT dengan optimasi pada proses *pretraining* menggunakan data yang lebih besar dan tanpa *Next Sentence Prediction*, sehingga dalam penelitian ini istilah BERT digunakan sebagai arsitektur payung yang mencakup RoBERTa sebagai pengembangannya. Secara arsitektural, RoBERTa-base memiliki 12 lapisan *Transformer encoder*, masing-masing lapisan terdiri dari 12 kepala *attention*, dimensi *hidden state* 768, dan total sekitar 125 juta parameter [19]. Representasi dari RoBERTa, baik dalam kondisi *pre-trained* maupun *fine-tuned*, terbukti menghasilkan fitur yang cukup kuat sehingga *classifier downstream* mampu mencapai performa yang kompetitif [20]. Arsitektur ekstraksi fitur menggunakan *dual pooling* pada *last_hidden_state*:

1. *Mean pooling*: rata-rata semua token non-*padding*
2. *Max pooling*: nilai maksimum per dimensi

Kedua vektor dikonkatendaskan menghasilkan vektor berdimensi 1536. Strategi *dual pooling* dengan mengonkatenasi *mean pooling* dan *max pooling* terbukti secara konsisten mengungguli penggunaan salah satu *pooling* saja sebagai representasi fitur untuk klasifikasi teks [20].

2.4.2. Ekstraksi Fitur untuk Task 1A

Untuk Task 1A (klasifikasi biner HOF vs. NOT), vektor 1536 dimensi hasil *dual pooling* langsung digunakan sebagai fitur, kemudian ditambahkan fitur L2-norm (1 dimensi) yang dihitung dari vektor *embedding*. Total fitur yang digunakan XGBoost pada Task 1A adalah 1537 dimensi. Pendekatan ini dipilih karena Task 1A merupakan klasifikasi biner dengan ruang keputusan yang relatif sederhana, sehingga representasi fitur penuh dari *dual pooling* dapat dipertahankan tanpa pengurangan dimensi. Penambahan L2-norm bertujuan untuk memberikan informasi skala vektor sebagai fitur tambahan yang bersifat global.

2.4.3. Ekstraksi Fitur untuk Task 1B

Untuk Task 1B (klasifikasi multi-kelas), *embedding* 1536 dimensi hasil *dual pooling* direduksi menggunakan *Principal Component Analysis* (PCA). Reduksi dimensi diperlukan karena ruang fitur 1536 dimensi berpotensi menimbulkan *curse of dimensionality* pada XGBoost karena distribusi kelas sangat tidak seimbang. PCA dipilih karena mampu mempertahankan variansi maksimum data dengan jumlah komponen yang jauh lebih kecil, sekaligus mengurangi korelasi antarfitur yang dapat mengganggu kinerja *tree-based classifier*. Jumlah komponen PCA sebesar 64 ditentukan melalui eksperimen *grid search* pada data validasi dengan menguji rentang komponen 32, 64, 128, dan 256. Konfigurasi 64 komponen menghasilkan *Macro F1* tertinggi dengan *explained variance* sekitar 85%, sehingga ditetapkan sebagai konfigurasi final. Kemudian ditambahkan:

1. *Soft probability* (4 dimensi): probabilitas keluaran *softmax* dari RoBERTa untuk setiap kelas.
2. *Entropy* (1 dimensi): ukuran ketidakpastian prediksi RoBERTa, dihitung sebagai $-\sum(p \cdot \log(p))$ dari *soft probability*.

Total fitur yang digunakan XGBoost pada Task 1B adalah 69 dimensi (64 + 4 + 1).

2.4.4. Fine-Tuning RoBERTa

Fine-tuning RoBERTa dilakukan secara berbeda untuk masing-masing *task* sesuai karakteristiknya. Untuk Task 1A (klasifikasi biner), *fine-tuning* menggunakan standar *Cross-Entropy Loss* dengan AdamW *optimizer* ($lr = 2 \times 10^{-5}$) selama 3 *epoch*, karena distribusi kelas yang relatif seimbang (rasio HOF:NOT $\approx 1,9:1$) tidak memerlukan penanganan *loss* khusus.

Untuk Task 1B, model di-*fine-tune* menggunakan *Focal Loss* untuk mengatasi ketidakseimbangan kelas yang lebih ekstrim. *Focal Loss* memodifikasi *Cross-Entropy Loss* dengan memberikan bobot lebih rendah pada sampel yang mudah diklasifikasikan, sehingga model lebih fokus pada kelas minoritas seperti HATE dan OFFN. *Focal Loss* didefinisikan pada persamaan 1.

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

dengan p_t adalah probabilitas model untuk kelas target, $\gamma = 3,0$ (parameter *focusing*), dan α_t adalah bobot kelas yang dihitung berdasarkan *inverse frequency* untuk mengatasi ketidakseimbangan frekuensi. Prinsip serupa diusulkan oleh Li et al. [21] dalam *Equalized Focal Loss* untuk data *long-tailed*, yang memodulasi *loss* per kategori berdasarkan tingkat ketidakseimbangan. Proses *fine-tuning* dilakukan maksimal 12 *epoch* dengan *early stopping* berdasarkan *Macro F1-Score* pada data validasi (*patience* = 3 *epoch*).

Arsitektur RoBERTa untuk Task 1B juga dilengkapi dengan lapisan *Dropout* ($p = 0,1$) sebelum *classifier head* untuk mengurangi *overfitting*.

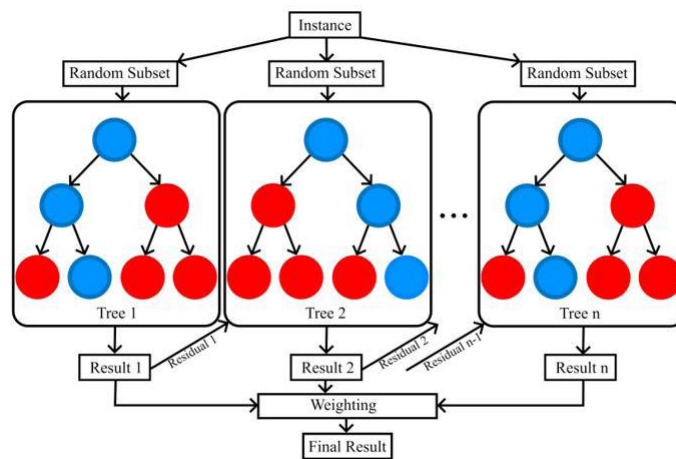
2.5. Balancing Data

Oversampling hanya diterapkan pada data training setelah proses *stratified split*, untuk menghindari *data leakage*. Implementasi *oversampling* dilakukan secara *custom* menggunakan fungsi manual berbasis NumPy, di mana indeks sampel kelas minoritas diduplikasi secara siklis menggunakan `np.resize` hingga jumlahnya sama dengan kelas mayoritas, kemudian seluruh data diacak urutannya sebelum proses pelatihan.

Pada Task 1A, kelas NOT di-oversampling dari 1.342 menjadi 2.501 (faktor $\sim 1,86\times$). Pada Task 1B, semua kelas minoritas (HATE, OFFN, PRFN) di-oversampling hingga menyamai jumlah kelas mayoritas NONE (1.342), dengan faktor duplikasi masing-masing $\sim 1,97\times$, $\sim 2,16\times$, dan $\sim 1,12\times$.

2.6. Pelatihan Model XGBoost

Setelah ekstraksi fitur dan *oversampling*, dilakukan pelatihan menggunakan algoritma XGBoost yang membangun model prediksi secara aditif melalui penjumlahan fungsi *decision tree* pada setiap iterasi. Gambar 2 mengilustrasikan struktur dasar XGBoost yang terdiri dari banyak pohon keputusan (*multiple trees*), di mana setiap pohon dilatih pada subset data yang berbeda (secara acak), dan hasil dari setiap pohon digabungkan melalui mekanisme pembobotan untuk menghasilkan prediksi akhir yang lebih stabil dan akurat.



Gambar 2. Struktur sederhana XGBoost yang terdiri dari banyak pohon Keputusan

Secara matematis, XGBoost membangun model prediksi secara aditif dengan menjumlahkan fungsi dari pohon keputusan pada setiap iterasi. Fungsi objektif yang dioptimasi terdiri dari fungsi *loss* dan komponen regularisasi yang digunakan untuk mengontrol kompleksitas model dapat dilihat pada persamaan 2.

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2)$$

dengan $l(\cdot)$ adalah fungsi *loss* (pada penelitian ini, multi:softmax untuk Task 1B dan binary:logistic untuk Task 1A), T adalah jumlah *leaf* pada pohon keputusan, dan w_j adalah bobot pada *leaf* ke- j . Parameter γ dan λ merupakan parameter regularisasi yang dioptimasi menggunakan Optuna pada rentang $\gamma = 0-5$ dan $\lambda = 10^{-8}-10$. XGBoost menggunakan pendekatan turunan orde kedua untuk mempercepat proses optimasi [22].

2.6.1. Baseline

Model *baseline* dilatih menggunakan BERT *pre-trained* (tanpa *fine-tuning*) sebagai *feature extractor* dengan mengambil token [CLS] (768 dimensi). XGBoost dikonfigurasi dengan parameter default (`n_estimators = 300`, `learning_rate = 0,1`, `max_depth = 6`) tanpa optimasi apa pun, sebagai batas bawah performa (*lower bound*).

2.6.2. XGBoost dengan Parameter Default (Predefined)

Setelah proses *fine-tuning* RoBERTa, ekstraksi fitur menggunakan *dual pooling* (*mean+max*) menghasilkan vektor 1536 dimensi, ditambah fitur L2-norm (1 dimensi) sehingga total 1537 fitur. Model XGBoost dilatih dengan konfigurasi parameter *default* (`n_estimators = 2000`, `max_depth = 6`, `learning_rate =`

0,05, disertai `early_stopping_rounds = 50`) tanpa optimasi tambahan. Tahap ini bertujuan untuk mengukur peningkatan performa yang berasal dari *fine-tuning* RoBERTa dan fitur yang lebih kaya, sebelum dilakukan optimasi *hyperparameter* lebih lanjut.

2.6.3. Optimasi Hyperparameter dengan Optuna

Optimasi *hyperparameter* menggunakan framework Optuna dengan algoritma TPE Sampler sebanyak 50 *trial*. Optuna merupakan framework optimasi *hyperparameter* generasi terbaru yang mengintegrasikan algoritma *Tree-structured Parzen Estimator* (TPE) untuk mengeksplorasi ruang pencarian secara efisien dengan memodelkan distribusi parameter yang menjanjikan dan tidak menjanjikan secara terpisah [23]. Pendekatan ini telah terbukti lebih efisien dibandingkan *grid search* maupun *random search* konvensional, terutama pada ruang *hyperparameter* berdimensi tinggi, sebagaimana divalidasi dalam berbagai studi optimasi arsitektur dan model [23], [24].

Hyperparameter yang dioptimasi meliputi `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `colsample_bylevel`, `min_child_weight`, `gamma`, `reg_alpha`, `reg_lambda`, dan `n_estimators`. *Objective function* adalah memaksimalkan *Macro F1-Score* pada data validasi.

2.6.4. Multi-Seed Ensemble

Sebanyak 5 model XGBoost dilatih dengan *hyperparameter* yang sama, namun menggunakan *random seed* berbeda, yaitu [42, 123, 456, 789, 2024]. Probabilitas keluaran dari kelima model dirata-ratakan (*soft voting*) untuk mengurangi varians prediksi. Mnassri *et al.* [25] menunjukkan bahwa penerapan teknik *ensemble*, seperti *soft voting* dan *stacking*, dapat meningkatkan kinerja deteksi *hate speech* pada data yang tidak seimbang.

2.7. Threshold

Threshold digunakan untuk menyesuaikan batas keputusan model guna meningkatkan performa pada data tidak seimbang. Pendekatan yang digunakan berbeda antara Task 1A dan Task 1B karena perbedaan sifat klasifikasi (biner vs multi-kelas):

1. Task 1A (biner): Cukup menggeser satu batas probabilitas (*threshold scan*).
2. Task 1B (multi-kelas): Digunakan metode *division calibration*, yaitu probabilitas setiap kelas dibagi dengan threshold per kelas, kemudian dipilih kelas dengan nilai tertinggi.

2.7.1. Threshold Scan (Khusus Task 1A)

Pada Task 1A (klasifikasi biner HOF vs NOT), prediksi akhir ditentukan dengan membandingkan probabilitas kelas HOF (p_{HOF}) dengan nilai threshold t yang dapat dilihat pada persamaan 3.

$$\hat{y} = \begin{cases} \text{HOF} & \text{jika } p_{HOF} > t \\ \text{NOT} & \text{sebaliknya} \end{cases} \quad (3)$$

Nilai t dicari melalui *threshold scan* pada data validasi, dengan rentang 0,10 hingga 0,91 dan *step* 0,01. Dipilih nilai t yang memaksimalkan *Macro F1-Score* pada data validasi.

2.7.2. Threshold Calibration (Khusus Task 1B)

Pada Task 1B (klasifikasi multikelas), metode *division calibration* diterapkan. Untuk setiap kelas c , probabilitas asli p_c dari model dibagi dengan threshold t_c yang dapat dilihat pada persamaan 4.

$$s_c = \frac{p_c}{t_c} \quad (4)$$

Kelas prediksi akhir dipilih berdasarkan nilai s_c tertinggi, yang dapat dilihat pada persamaan 5.

$$\hat{y} = \arg \max_c s_c \quad (5)$$

Threshold untuk NONE dan PRFN ditetapkan sebesar 0,5 karena distribusi probabilitasnya sudah stabil. Threshold untuk kelas minoritas HATE dan OFFN dioptimasi dengan *grid search* pada rentang 0,3–0,7 (15 nilai *step*) untuk memaksimalkan *Macro F1-Score* pada data validasi.

Strategi penyesuaian *threshold* (*decision threshold calibration*) merupakan pendekatan *post-processing* yang efektif untuk menangani ketidakseimbangan kelas tanpa pelatihan ulang model [26]. Penelitian terkini menunjukkan bahwa *threshold calibration* mampu meningkatkan performa klasifikasi pada data tidak seimbang dan dalam beberapa kasus memberikan hasil yang lebih konsisten dibandingkan dengan teknik *resampling* konvensional [27]. Dengan menurunkan *threshold* pada kelas minoritas, *recall* kelas tersebut meningkat, sehingga berkontribusi pada peningkatan *Macro F1-Score*.

2.8. Metrik Evaluasi

Metrik utama yang digunakan adalah Macro F1-Score, yang menghitung rata-rata F1-Score dari seluruh kelas dengan bobot yang sama, sehingga lebih representatif untuk dataset yang tidak seimbang. F1-Score per kelas dihitung menggunakan persamaan 6, kemudian dirata-ratakan menjadi Macro F1-Score pada persamaan 6. Untuk Task 1A (klasifikasi biner, kelas NOT dan HOF):

$$Macro-F1_{1A} = \frac{F1_{NOT} + F1_{HOF}}{2} \quad (6)$$

Untuk Task 1B (klasifikasi multi-kelas, kelas NONE, HATE, OFFN, PRFN):

$$Macro-F1_{1B} = \frac{F1_{NONE} + F1_{HATE} + F1_{OFFN} + F1_{PRFN}}{4} \quad (7)$$

Dengan pendekatan ini, setiap kelas berkontribusi sama, sehingga performa pada kelas minoritas (HATE dan OFFN) tidak terabaikan.

3. HASIL DAN PEMBAHASAN

3.1. Hasil

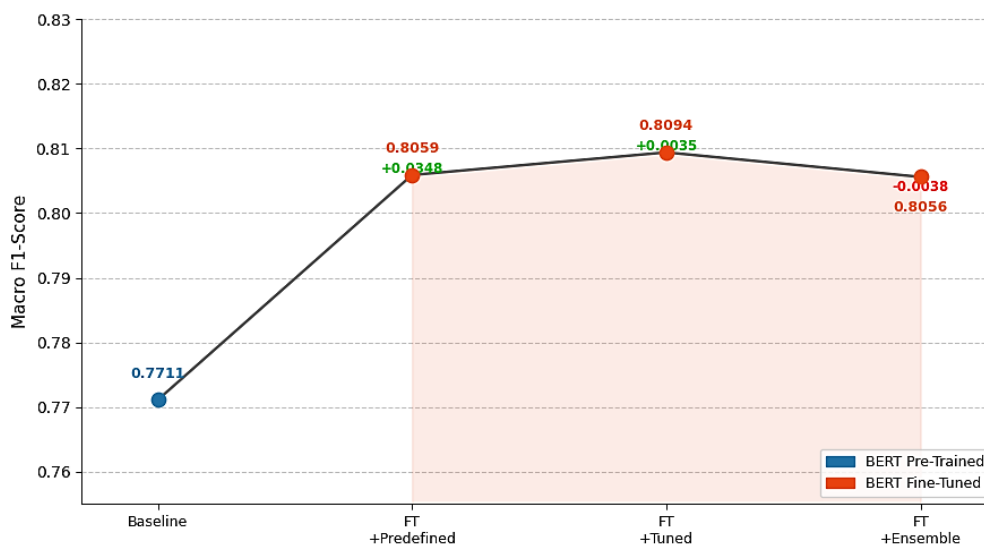
3.1.1. Perbandingan Performa Validasi Task 1A

Tabel 3 menyajikan hasil evaluasi seluruh konfigurasi model pada data validasi (20% *hold out*) untuk Task 1A. Model RoBERTa *Fine-Tuned* + XGBoost (*Tuned*) dengan *threshold* 0,55 mencapai *Macro F1-Score* tertinggi sebesar 0,8094, mengungguli RoBERTa *Pre-Trained* + XGBoost (0,7711). Peningkatan dari RoBERTa *Pre-Trained* ke RoBERTa *Fine-Tuned* mencapai 3,83 poin persentase, menunjukkan manfaat *fine-tuning* pada data target.

Tabel 3. Hasil Validasi Task 1A (Macro F1-Score)

Model	Konfigurasi	Fitur	Macro F1-Score	Threshold
RoBERTa Pre-Trained + XGB	<i>Baseline</i>	768 (CLS token)	0,7711	-
RoBERTa Fine-Tuned + XGB	<i>Predefined</i>	1536 + L2-norm	0,8059	0,45
RoBERTa Fine-Tuned + XGB	<i>Tuned</i>	1536 + L2-norm	0,8094	0,55
RoBERTa Fine-Tuned + XGB	<i>Ensemble</i>	1536 + L2-norm	0,8056	0,56

(FT = *Fine-Tuned*; *Predefined* = parameter default; *Tuned* = hasil Optuna; *Ensemble* = rata-rata 5 seed)



Gambar 3. Grafik Macro F1-Score Model HASOC 2021 Task 1A

Pada Task 1A, model dengan optimasi Optuna (*Tuned*) mencapai *Macro F1-Score* tertinggi (0,8094), diikuti *predefined* (0,8059) dan *ensemble* (0,8056). Hal ini menunjukkan bahwa untuk klasifikasi biner, *tuning hyperparameter* lebih krusial daripada *ensemble*.

3.1.2. Perbandingan Performa Validasi Task 1B

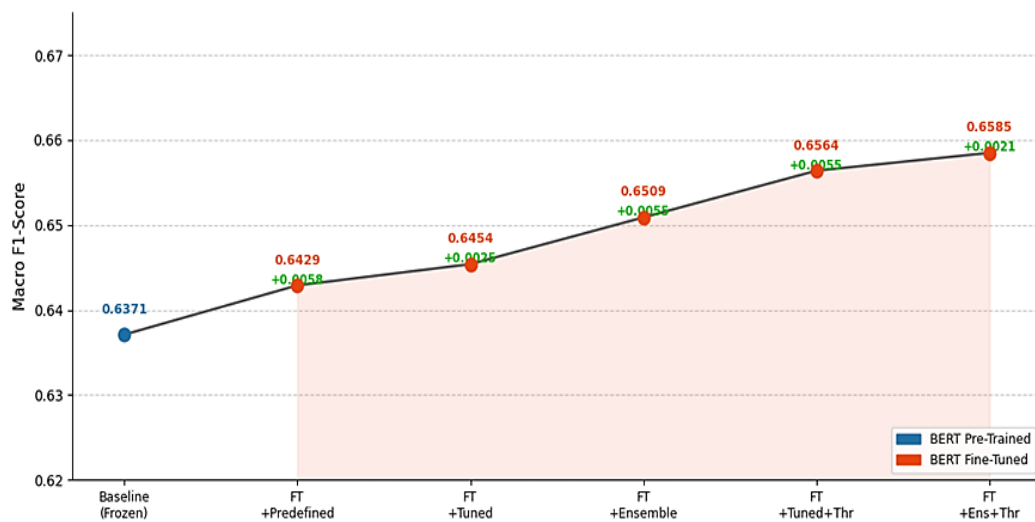
Tabel 4 menyajikan hasil evaluasi Task 1B. Model RoBERTa *Fine-Tuned* + XGBoost (*Ensemble+Threshold*) mencapai *Macro F1-Score* tertinggi sebesar 0,6585 dengan *threshold* [0,5; 0,414; 0,7; 0,5] untuk kelas [NONE; HATE; OFFN; PRFN]. RoBERTa *Pre-Trained* + XGBoost hanya mencapai 0,6371

(menggunakan *embedding* [CLS] 768 dimensi), menegaskan pentingnya *fine-tuning* untuk masalah multi-kelas yang lebih kompleks. Peningkatan dari *baseline pre-trained* ke model terbaik *fine-tuned* mencapai +0,0214 poin persentase.

Tabel 4. Hasil Validasi Task 1B (Macro F1-Score)

Model	Konfigurasi	Fitur	Macro F1-Score	Threshold
RoBERTa Pre-Trained + XGB	<i>Baseline</i>	768 (CLS token)	0,6371	-
RoBERTa Fine-Tuned + XGB	Predefined	PCA64+Prob+Entropy	0,6429	-
RoBERTa Fine-Tuned + XGB	Tuned (Optuna)	PCA64+Prob+Entropy	0,6454	-
RoBERTa Fine-Tuned + XGB	Ensemble (5 seeds)	PCA64+Prob+Entropy	0,6509	-
RoBERTa Fine-Tuned + XGB	Tuned+Threshold	PCA64+Prob+Entropy	0,6564	[0,5; 0,386; 0,643; 0,5]
RoBERTa Fine-Tuned + XGB	Ensemble+Threshold	PCA64+Prob+Entropy	0,6585	[0,5; 0,414; 0,7; 0,5]

(FT = Fine-Tuned; Predefined = parameter default; Tuned = hasil Optuna; Ensemble = rata-rata 5 seed)



Gambar 4. Grafik Macro F1-Score Model HASOC 2021 Task 1B

Pada Task 1B, konfigurasi *ensemble + threshold calibration* memberikan peningkatan performa paling signifikan, dengan *Macro F1-Score* 0,6585. Angka ini lebih tinggi +2,14 poin persentase dibandingkan *baseline* (0,6371) dan juga melampaui *ensemble* tanpa kalibrasi (0,6509).

3.2. Hasil Pengujian pada Data Uji (Test Set)

Tabel 5 menyajikan hasil prediksi pada data uji HASOC 2021 (1.281 *tweet*) dengan model terbaik. Nilai *Macro F1-Score* pada data uji diperoleh dengan membandingkan prediksi model terhadap *ground truth* yang disediakan oleh penyelenggara kompetisi HASOC 2021.

Tabel 5. Hasil Prediksi pada Data Uji

Task	Model Terbaik	Macro F1-Score	Prediksi Kelas					
			NONE	HATE	OFFN	PRFN	HOF	NOT
Task 1A	RoBERTa FT + XGB (<i>Tuned</i>)	0,8083	—	—	—	—	795	486
Task 1B	RoBERTa FT + XGB (<i>Ens+Thr</i>)	0,6541	418	267	173	423	—	—

(FT = Fine-Tuned)

3.3. Analisis Perbandingan dengan Penelitian Sebelumnya

Tabel 6 menyajikan perbandingan hasil penelitian ini dengan tim-tim dari kompetisi HASOC 2021 yang menggunakan pendekatan RoBERTa, XGBoost, atau keduanya. Seluruh nilai pada kolom Tim Lain

merupakan *Macro F1-Score* resmi dari papan peringkat HASOC 2021 (data uji). Nilai untuk *Proposed Method* adalah hasil evaluasi pada data uji yang sama. Penelitian ini mencapai *Macro F1-Score* 0,8083 pada Task 1A dan 0,6541 pada Task 1B (data uji). Jika disandingkan dengan papan peringkat resmi, hasil Task 1A akan menempati posisi ke-5, sementara Task 1B berada di peringkat ke-3, hanya terpaut 0,0036 dari peringkat kedua (0,6577).

Tabel 6. Perbandingan dengan Penelitian Sebelumnya

No.	Tim / Metode	Model	Task 1A (HOF vs NOT)		Task 1B (NONE/HATE/OFFN/PR FN)	
			Macro F1-Score	Rank	Macro F1-Score	Rank
1	HUNLP [28]	Text GCN dengan GloVe Twitter embedding + TF-IDF & PMI	0,8215	2	0,6296	9
2	neuro utmn thales [29]	Soft-voting ensemble dari 5 Twitter-RoBERTa fine-tuned + HatebaseTwitter	0,8199	3	0,6577	2
3	HNLP [30]	RoBERTa + Modified Cross Entropy Loss	0,8089	4	0,6396	6
4	UINSUSKA [31]	BERT-base-uncased + Feed Forward Neural Network	0,8024	8	0,6417	5
5	NeuralSpace [32]	Multilingual XLM-RoBERTa + emoji2vec	0,7996	14	0,6268	11
6	IRLab@IITBHU [33]	Fine-tuned BERT-Large-Cased	0,7976	16	0,6093	18
7	Vishesh Gupta [10]	GloVe+GRU (1A), TF-IDF+XGBoost (1B)	0,7680	27	0,5871	22
8	Proposed Method	RoBERTa Fine-Tuned + XGBoost (Tuned + Ensemble + Threshold)	0,8083	-	0,6541	-

Catatan: Nilai *Proposed Method* adalah hasil evaluasi pada data uji (test set) HASOC 2021 yang diberikan oleh penguji. Seluruh nilai tim lain bersumber dari papan peringkat resmi kompetisi HASOC 2021 (Mandl et al. [9]).

Tidak ada tim HASOC 2021 yang menggabungkan RoBERTa dengan XGBoost; satu-satunya tim yang menggunakan XGBoost [10] justru menggunakan TF-IDF dan menghasilkan skor lebih rendah. Hasil ini menegaskan bahwa pendekatan hybrid RoBERTa+XGBoost kompetitif dengan Transformer murni.

3.4. Pembahasan

Hasil eksperimen menunjukkan bahwa *fine-tuning* RoBERTa secara konsisten meningkatkan performa dibandingkan dengan RoBERTa *pre-trained*; peningkatan mencapai +3,83 poin persentase pada Task 1A (dari 0,7711 menjadi 0,8094) dan +2,14 poin persentase pada Task 1B (dari 0,6371 menjadi 0,6585). Pada Task 1A, konfigurasi *Tuned* (Optuna) menghasilkan *Macro F1* tertinggi (0,8094), sedikit di atas *predefined* (0,8059) dan *ensemble* (0,8056), yang mengindikasikan bahwa optimasi *hyperparameter* lebih krusial daripada *ensemble* untuk klasifikasi biner. Pada Task 1B, *threshold calibration* berhasil menurunkan *threshold* kelas HATE dari 0,5 menjadi 0,414 dan menaikkan *threshold* OFFN menjadi 0,7, sehingga *recall* HATE meningkat dari 0,56 menjadi 0,64 tanpa mengorbankan *precision*. Dibandingkan dengan peserta kompetisi HASOC 2021, metode yang diusulkan mencapai posisi setara dengan peringkat ke-5 (Task 1A) dan ke-3 (Task 1B) di antara pendekatan berbasis RoBERTa. Selain itu, *oversampling* pada ruang fitur terbukti efektif, terutama setelah reduksi dimensi menjadi 69 fitur.

3.4.1. Keunggulan

Arsitektur hybrid RoBERTa+XGBoost unggul dalam fleksibilitas optimasi. Konfigurasi *Tuned* (Optuna) mencapai *Macro F1* tertinggi pada Task 1A, sedangkan *ensemble* dengan *threshold calibration* memberikan peningkatan paling signifikan pada Task 1B. *Threshold calibration* berhasil meningkatkan *recall* kelas minoritas HATE tanpa mengorbankan *precision*. Model juga menunjukkan kemampuan generalisasi yang baik dengan selisih *Macro F1* antara validasi dan uji yang sangat kecil (0,0011 untuk Task 1A, 0,0044 untuk Task 1B).

3.4.2. Keterbatasan

Kelas OFFN menjadi yang paling sulit diklasifikasikan, dengan F1-Score hanya 0,50, yang mengindikasikan tumpang tindih semantik yang tinggi dengan kategori HATE dan PRFN. Penggunaan PCA menyisakan sekitar 15% variansi data yang tidak terjelaskan (explained variance 85%), sehingga berpotensi menghilangkan informasi halus. Model hanya diuji pada dataset HASOC 2021 berbahasa Inggris, sehingga generalisasi ke bahasa lain belum terjamin.

3.4.3. Analisis Kesalahan

Analisis kesalahan pada data validasi Task 1B menunjukkan bahwa kelas OFFN merupakan kelas yang paling sulit diklasifikasikan (F1-Score = 0,50), dengan banyak tweet OFFN yang salah diprediksi sebagai HATE atau PRFN akibat tumpang tindih semantik yang tinggi. Model cenderung mengasosiasikan bahasa keras dengan ujaran kebencian terhadap kelompok (HATE), padahal perbedaan kunci antara OFFN dan HATE terletak pada target ujaran, yaitu individu versus kelompok, yang sulit ditangkap dari representasi embedding saja. Kelas HATE sendiri menunjukkan banyak false positive (precision = 0,58) karena tweet dari kelas lain yang mengandung bahasa emosional kerap salah dikategorikan sebagai HATE. Sebaliknya, kelas PRFN dan NONE relatif lebih mudah dikenali karena memiliki pola leksikal yang lebih konsisten.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi *hate speech* dan *offensive language* menggunakan pendekatan hybrid *Robustly Optimised BERT Approach* (RoBERTa) + XGBoost pada dataset *Hate Speech and Offensive Content Identification* (HASOC) 2021. Model RoBERTa *Fine-Tuned* + XGBoost (*Tuned*) mencapai *Macro F1-Score* 0,8083 pada Task 1A (peringkat ke-5 di antara seluruh peserta HASOC 2021), sedangkan model RoBERTa *Fine-Tuned* + XGBoost (*Ensemble+Threshold*) mencapai 0,6541 pada Task 1B (peringkat ke-3 di antara pendekatan berbasis RoBERTa). Kombinasi yang diusulkan tidak ditemukan dalam laporan peserta HASOC 2021 dan mampu bersaing secara kompetitif dengan pendekatan Transformer murni.

Teknik *oversampling* pada ruang fitur terbukti efektif dalam menangani ketidakseimbangan kelas. Optimasi *hyperparameter* dengan Optuna, *multi-seed ensemble*, *threshold scan* (Task 1A), dan *threshold calibration* (Task 1B) memberikan peningkatan performa yang saling melengkapi. Model menunjukkan kemampuan generalisasi yang baik dengan selisih *Macro F1-Score* antara validasi dan data uji yang sangat kecil (0,0011 untuk Task 1A, 0,0044 untuk Task 1B).

Penelitian ini memiliki beberapa limitasi. Model hanya diuji pada dataset HASOC 2021 berbahasa Inggris, sehingga generalisasi ke bahasa lain belum terjamin. Reduksi *Principal Component Analysis* (PCA) menyisakan sekitar 15% variansi data, sehingga berpotensi menghilangkan informasi halus. Kelas *Offensive Language* (OFFN) masih sulit dikenali (F1=0,50) karena tumpang tindih semantik dengan *Hate Speech* (HATE) dan *Profane Language* (PRFN). Untuk penelitian selanjutnya, disarankan menguji model pada dataset multibahasa, mengeksplorasi metode reduksi dimensi alternatif, serta menambahkan fitur berbasis entitas untuk membedakan ujaran kebencian terhadap individu dari ujaran kebencian terhadap kelompok.

REFERENSI

- [1] A. Guterres, "Uncontrolled Social Media Networks Amplify The Worst Human Impulses, Allowing Hate Speech To Spread Like Wildfire And Ignite Violence," Dawn. Accessed: Apr. 17, 2026. [Online]. Available: <https://www.dawn.com/news/1874996>
- [2] X Corp., "Global Transparency Report H2 2024." Accessed: Jun. 13, 2026. [Online]. Available: https://transparency.x.com/content/dam/transparency-twitter/2025/x-global-transparency-report_h2_2024.pdf
- [3] X Corp., "Global Transparency Report H1 2024." Accessed: Jun. 13, 2026. [Online]. Available: <https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf>
- [4] P. Madriaza et al., "Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities," Mar. 01, 2025, *John Wiley and Sons Inc.* doi: 10.1002/cl2.70018.
- [5] University of Oxford, "Majority support moderation on social media platforms, global survey shows," University of Oxford. Accessed: Apr. 17, 2026. [Online]. Available: <https://www.ox.ac.uk/news/2025-02-12-majority-support-moderation-social-media-platforms-global-survey-shows>
- [6] Republik Indonesia, *Undang-Undang Republik Indonesia Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik sebagaimana telah diubah dengan Undang-Undang Nomor 19 Tahun 2016*. Indonesia: Lembaran Negara RI Tahun 2016 No. 251, 2016.
- [7] T. Mandl et al., "Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages." [Online]. Available: <https://doi.org/10.1145/3368567.3368584>

- [8] T. Mandl, S. Modha, G. K. Shahi, and others, "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages," in *CEUR Workshop Proc. (FIRE 2020)*, 2020. doi: 10.1145/3441501.3441517.
- [9] T. Mandl, G. K. Shahi, M. Zampieri, and others, "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages," in *FIRE 2021 (CEUR Workshop Proceedings)*, 2021. doi: 10.48550/arXiv.2112.09301.
- [10] R. Kumar, V. Gupta, and R. Pamula, "Hate Speech and Offensive Content Identification in English Tweets," 2021. [Online]. Available: <https://ceur-ws.org/Vol-3159/T1-46.pdf>
- [11] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model," *PLoS One*, vol. 15, no. 8, p. e0237861, 2020, doi: 10.1371/journal.pone.0237861.
- [12] R. Sivanaiyah, A. D. S. S. M. Rajendram, and T. T. Mirnalinee, "TECHSSN at SemEval-2020 Task 12: Offensive Language Detection Using BERT Embeddings," in *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, 2020. doi: 10.18653/v1/2020.semeval-1.291.
- [13] M. Almaliki, A. M. Almars, I. Gad, and E. S. Atlam, "ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media," *Electronics (Switzerland)*, vol. 12, no. 4, Feb. 2023, doi: 10.3390/electronics12041048.
- [14] J. Salminen, M. Hopf, S. A. Chowdhury, S. gyo Jung, H. Almerexhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences 2020 10:1*, vol. 10, no. 1, pp. 1-, Jan. 2020, doi: 10.1186/S13673-019-0205-6.
- [15] U. Iftikhar, S. F. Ali, G. Mustafa, N. Bahar, and K. Ishaq, "Beyond words: a hybrid transformer-ensemble approach for detecting hate speech and offensive language on social media," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.3214.
- [16] W. Ashiq *et al.*, "Roman urdu hate speech detection using hybrid machine learning models and hyperparameter optimization," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-79106-7.
- [17] CardiffNLP, "twitter-roberta-base-offensive." Accessed: Apr. 16, 2026. [Online]. Available: <https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive>
- [18] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," Oct. 2020, doi: 10.48550/arXiv.2010.12421.
- [19] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, doi: 10.48550/arXiv.1907.11692.
- [20] S. Alsaidi, M. Vincent, O. Boyer, N. Garcelon, M. Couceiro, and A. Coulet, "Comparing representations of long clinical texts for the task of patient note-identification," Mar. 2025, doi: 10.18653/v1/2025.cl4health-1.9.
- [21] B. Li *et al.*, "Equalized Focal Loss for Dense Long-Tailed Object Detection," Jun. 2022, doi: 10.1109/CVPR52688.2022.00686.
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019, doi: 10.1145/3292500.3330701.
- [24] M. Sipper, "Neural Networks with À La Carte Selection of Activation Functions," *SN Comput. Sci.*, vol. 2, no. 6, Nov. 2021, doi: 10.1007/s42979-021-00885-1.
- [25] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT-based Ensemble Approaches for Hate Speech Detection," in *Proceedings - IEEE Global Communications Conference, GLOBECOM*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 4649–4654. doi: 10.1109/GLOBECOM48099.2022.10001325.
- [26] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning," Jun. 28, 2021, *American Chemical Society*. doi: 10.1021/acs.jcim.1c00160.
- [27] M. Abdelhamid and A. Desai, "Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification," Sep. 2024, doi: 10.48550/arXiv.2409.19751.
- [28] N. Bölücü and P. Canbay, "Hate Speech and Offensive Content Identification with Graph Convolutional Networks," 2021. [Online]. Available: <https://ceur-ws.org/Vol-3159/T1-4.pdf>
- [29] A. Glazkova, M. Kadantsev, and M. Glazkov, "Fine-tuning of Pre-trained Transformers for Hate, Offensive, and Profane Content Detection in English and Marathi," Oct. 2021, doi: 10.48550/arXiv.2110.12687.
- [30] A. Mitra and P. Sankhala, "Multilingual Hate Speech and Offensive Content Detection using Modified Cross-entropy Loss," Feb. 2022, doi: 10.48550/arXiv.2202.02635.

- [31] S. Agustian, R. Saputra, and A. Fadhilah, “‘Feature Selection’ With Pretrained-Bert For Hate Speech And Offensive Content Identification In English And Hindi Languages,” 2021. doi: 10.13140/RG.2.2.27831.60324.
- [32] M. Bhatia *et al.*, “One to rule them all: Towards Joint Indic Language Hate Speech Detection,” Sep. 2021, doi: 10.48550/arXiv.2109.13711.
- [33] S. Chanda, S. Ujjwal, S. Das, and S. Pal, “Fine-tuning Pre-Trained Transformer based model for Hate Speech and Offensive Content Identification in English, Indo-Aryan and Code-Mixed (English-Hindi) languages,” 2021. [Online]. Available: <https://ceur-ws.org/Vol-3159/T1-44.pdf>