Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

Klasifikasi Ujaran Kebencian Menggunakan 5-Fold Ensemble Weighted Probability Averaging berbasis Arsitektur Twitter-RoBERTa

Ibra Sahrian Alsa, Surya Agustian*, Fitra Kurnia, Pizaini, Siska Kurnia Gusti

Prodi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12150114195@students.uin-suska.ac.id, [2,*]surya.agustian@uin-suska.ac.id, [3]fitra.k@uin-suska.ac.id, [4]pizaini@uin-suska.ac.id, [5]siskakurniagusti@uin-suska.ac.id
Email Penulis Korespondensi: surya.agustian@uin-suska.ac.id

Abstrak−Penyebaran ujaran kebencian di media sosial memicu dampak psikologis negatif dan polarisasi sosial. Deteksi otomatis konten toksik ini menghadapi tantangan ambiguitas linguistik, pergeseran distribusi data secara temporal, dan instabilitas fine-tuning Transformer pada dataset kecil yang tidak seimbang. Penelitian ini mengusulkan arsitektur end-to-end yang mengintegrasikan model Twitter-RoBERTa dengan teknik ensemble learning. Metode komputasional yang diimplementasikan mencakup Bayesian Optimization melalui framework Optuna untuk otomatisasi pencarian hyperparameter, pemisahan data dengan Stratified 5-Fold Cross-Validation, serta penggabungan keputusan berbasis probabilitas menggunakan algoritma Weighted Probability Averaging (WPA) sebagai kontribusi kebaruan arsitektural. Tujuan utamanya adalah menstabilkan pelatihan dan mengatasi bias dominasi kelas mayoritas dengan memberikan porsi keputusan lebih besar kepada model fold berkinerja tertinggi. Evaluasi pada korpus gabungan HASOC 2020 dan 2021 menunjukkan bahwa arsitektur ini mencapai Macro F1-Score sebesar 80,99% untuk klasifikasi biner (Subtask 1A) dan 64,70% untuk klasifikasi fine-grained multi-kelas (Subtask 1B). Capaian ini membuktikan metode ensemble WPA sukses mengungguli pendekatan model tunggal konvensional dan menempatkan sistem ini pada posisi top-4 global papan peringkat resmi HASOC 2021.

Kata Kunci: Deteksi Ujaran Kebencian; Twitter-RoBERTa; 5-Fold Ensemble; Weighted Probability Averaging; Bayesian Optimization

Abstract−*The proliferation of hate speech on social media triggers negative psychological impacts and social polarization. Automatic detection of this toxic content faces significant challenges, including linguistic ambiguity, temporal data distribution shifts, and Transformer fine-tuning instability on small, imbalanced datasets. To address these issues, this study proposes an end-to-end architecture that integrates the Twitter-RoBERTa model with ensemble learning techniques. The implemented computational methods include Bayesian Optimization via the Optuna framework for automated hyperparameter search, data partitioning using Stratified 5-Fold Cross-Validation, and probability-based decision fusion using the Weighted Probability Averaging (WPA) algorithm as an architectural novelty. The primary objective is to stabilize the fine-tuning process and mitigate majority class dominance bias by dynamically assigning a higher decision weight to the highest-performing fold models. Evaluation on the combined HASOC 2020 and 2021 corpora demonstrates that the proposed architecture achieves a Macro F1-Score of 80.99% for binary classification (Subtask 1A) and 64.70% for multi-class fine-grained classification (Subtask 1B). These results confirm that the WPA ensemble method successfully outperforms conventional single-model approaches, placing the system in a top-4 global position on the official HASOC 2021 leaderboard.*

Keywords: Hate Speech Detection; Twitter-RoBERTa; 5-Fold Ensemble; Weighted Probability Averaging; Bayesian Optimization

1. PENDAHULUAN

Platform media sosial secara fundamental telah mengubah paradigma komunikasi dan pertukaran informasi global dengan melintasi batas geografis dan budaya. Kendati demikian, kebebasan berekspresi pada platform tersebut sering kali disalahgunakan sebagai sarana utama untuk menyebarkan ujaran kebencian (*hate speech*), narasi yang merendahkan martabat, serta provokasi kekerasan yang menargetkan kelompok minoritas. Berdasarkan Laporan Transparansi Global X Corporation pada paruh pertama tahun 2024, platform tersebut menerima lebih dari 66,8 juta laporan pengguna terkait perilaku kebencian (*hateful conduct*), yang berimplikasi pada tindakan penegakan terhadap jutaan unggahan [1]. Tingginya volume konten berbahaya tersebut menyebabkan mekanisme moderasi konten secara manual menjadi tidak efisien secara ekonomi, membutuhkan waktu yang lama, serta berisiko memicu trauma psikologis berupa *Secondary Traumatic Stress* bagi para moderator [2]. Oleh karena itu, pengembangan sistem deteksi ujaran kebencian otomatis berbasis *Natural Language Processing* (NLP) memiliki urgensi yang tinggi untuk menjaga keamanan ekosistem digital. Tantangan utama dalam membangun sistem otomatis ini terletak pada tingginya ambiguitas linguistik, mengingat teks ujaran kebencian di media sosial kerap memuat sarkasme, penggunaan ejaan tidak baku, dialek spesifik seperti *African American English* (AAE), serta bahasa bersandi yang sangat bergantung pada konteks semantik .

Secara historis, sistem deteksi awal ujaran kebencian umumnya mengandalkan pendekatan *machine learning* tradisional seperti *Support Vector Machine* (SVM) atau *Naive Bayes* dengan metode ekstraksi fitur *Bag-of-Words* (BoW) atau *Term Frequency-Inverse Document Frequency* (TF-IDF) [3]. Meskipun efisien dari segi komputasi, metode klasik tersebut rentan terhadap misklasifikasi karena hanya menitikberatkan pada pencocokan kata kunci secara harfiah. Akibatnya, sistem sering menghasilkan nilai *false positive* pada kalimat yang memuat kosakata kasar namun tidak mengandung intensi ujaran kebencian, seperti pada konteks *counter-speech* atau humor. Dalam perkembangannya, teknologi *state-of-the-art* (SOTA) di bidang *Natural Language Processing* (NLP) kini didominasi oleh arsitektur Transformer, khususnya *Bidirectional Encoder Representations from Transformers* (BERT) serta variannya yang dioptimasi secara *robust*, yaitu RoBERTa. Model-model tersebut mengintegrasikan mekanisme *self-attention* untuk menangkap hubungan kontekstual antar-kata secara dua arah (*bidirectional*). Varian spesifik domain seperti *Twitter-RoBERTa* bahkan telah dilatih menggunakan jutaan data cuitan, sehingga menjadi arsitektur *baseline* yang kuat untuk

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

menangani karakteristik unik teks media sosial [4]. Kendati arsitektur Transformer mampu memberikan peningkatan akurasi yang signifikan, penerapannya untuk klasifikasi ujaran kebencian pada kondisi riil masih terhambat oleh beberapa kelemahan fundamental dalam metodologi penelitian terdahulu yang kemudian menjadi *research gap* dalam studi ini.

Kesenjangan penelitian pertama terletak pada kerentanan model terhadap *distributional shift*. Beberapa penelitian terkait, seperti yang dilakukan oleh Antypas & Camacho-Collados (2023), menunjukkan bahwa model deteksi ujaran kebencian rentan mengalami penurunan performa yang signifikan ketika dievaluasi pada rentang waktu yang berbeda karena karakteristik dan topik ujaran kebencian terus berevolusi [5]. Model konvensional yang sangat bergantung pada *dataset* statis rentan mengalami *in-distribution overfitting*. Kondisi ini menyebabkan model cenderung hanya menghafal kata kunci atau tagar spesifik dari suatu peristiwa tertentu dan gagal menggeneralisasi fitur semantik ketika diuji pada data dari lini waktu yang berbeda [6]. Penggabungan korpus lintas tahun secara langsung juga terbukti menyuntikkan *noise* distribusional yang tidak mampu disaring secara memadai oleh model Transformer tunggal.

Kesenjangan kedua berkaitan dengan instabilitas optimasi saat melakukan *fine-tuning* arsitektur Transformer pada *dataset* berskala kecil hingga menengah. Penelitian terkait oleh Mosbach et al. (2021) dan Dodge et al. (2020) menemukan bahwa proses *fine-tuning* model Transformer memiliki sensitivitas yang tinggi terhadap inisialisasi *random seed* dan pemilihan *hyperparameter* [7], [8]. Model dengan arsitektur identik yang dilatih menggunakan konfigurasi serupa namun dengan nilai *seed* yang berbeda dapat menghasilkan variansi metrik performa yang lebar akibat fenomena *vanishing gradient* pada lapisan bawah Transformer. Kendati demikian, mayoritas penelitian terdahulu kerap mengabaikan masalah instabilitas tersebut atau hanya mengandalkan metode pencarian konvensional seperti *Grid Search* dan *Random Search*. Metode *Grid Search* mengevaluasi seluruh kombinasi parameter secara menyeluruh sehingga membutuhkan sumber daya komputasi yang besar secara eksponensial, sedangkan *Random Search* menguji nilai secara acak tanpa memanfaatkan riwayat evaluasi sebelumnya [6]. Akibatnya, kedua metode pembanding tersebut rentan gagal menemukan titik konvergensi global yang optimal.

Kesenjangan penelitian ketiga terdapat pada pendekatan *ensemble learning* yang digunakan untuk meningkatkan ketahanan model. Penelitian terkait oleh Kucukkaya & Toraman (2024) dan Yoo et al. (2025) membuktikan bahwa penggabungan beberapa model Transformer melalui teknik *ensemble* dapat meningkatkan akurasi secara signifikan [1], [9]. Kendati demikian, keterbatasan dari penelitian-penelitian tersebut terletak pada ketergantungannya terhadap metode agregasi konvensional, seperti *Majority Voting* atau *Simple Averaging* [1]. Dalam kasus deteksi ujaran kebencian, *dataset* yang digunakan umumnya mengalami ketidakseimbangan kelas yang tinggi, dengan kelas ujaran kebencian sebagai kelas minoritas. Pendekatan *ensemble* konvensional ini kurang optimal dalam mengklasifikasikan kelas minoritas karena nilai probabilitas keputusan akhir cenderung memihak pada model yang bias terhadap kelas mayoritas.

Untuk mengatasi ketiga *research gap* tersebut, penelitian ini mengusulkan, mengimplementasikan, dan mengevaluasi sebuah *pipeline* NLP *end-to-end* yang mengintegrasikan tiga intervensi metodologis secara sinergis pada arsitektur *Twitter-RoBERTa*. Pertama, penelitian ini mengatasi masalah instabilitas *fine-tuning* dengan menerapkan *Bayesian Optimization* melalui *framework* Optuna [10]. Berbeda dengan metode *Grid Search*, optimasi Bayesian mengonstruksi model probabilitas *surrogate* yang memanfaatkan riwayat eksperimen sebelumnya untuk memandu pencarian *learning rate* dan *batch size* secara terarah dan adaptif. Kedua, untuk memitigasi bias inisialisasi acak dan pergeseran distribusi, digunakan partisi *Stratified 5-Fold Cross-Validation* untuk melatih lima model secara independen pada porsi data yang saling melengkapi. Ketiga, sebagai bentuk kebaruan arsitektural, penelitian ini menggunakan metode *Weighted Probability Averaging* (WPA) sebagai pengganti agregasi konvensional. Agregasi WPA bertujuan untuk memberikan bobot yang dikalibrasi secara dinamis, model *fold* yang secara empiris memiliki *F1-Score* validasi lebih tinggi akan diberikan kontribusi keputusan yang lebih besar, sehingga secara efektif menyeimbangkan estimasi probabilitas pada kelas data minoritas. Seluruh *pipeline* ini dievaluasi secara komprehensif menggunakan *dataset* gabungan dari kompetisi *Hate Speech and Offensive Content Identification* (HASOC) edisi 2020 dan 2021 yang sarat akan *noise* distribusional. Melalui intervensi ini, penelitian ini diharapkan dapat menghasilkan arsitektur deteksi ujaran kebencian yang akurat, stabil, *robust* terhadap variasi temporal, serta kompetitif dalam standar metrik evaluasi global.

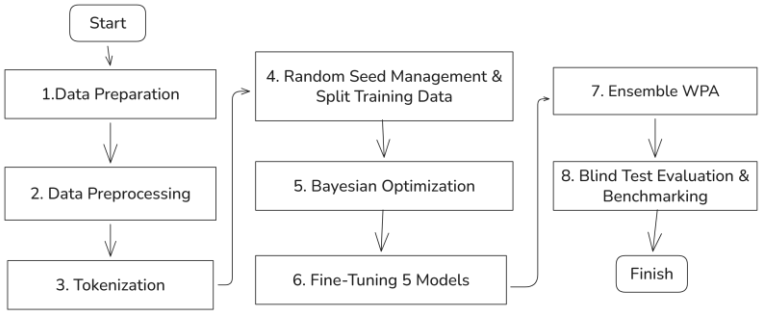
2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini menggunakan Pipeline NLP End-to-End yang terstruktur sebagai desain eksperimen progresif. Gambar 1 mengilustrasikan alur penelitian lengkap, yang mencakup tujuh tahapan operasional berbeda mulai dari persiapan data hingga benchmarking akhir terhadap leaderboard resmi HASOC 2021.

Pipeline ini mengikuti kerangka kerja sequential feedback loop yang dirancang secara terukur, di mana setiap tahap hulu menghasilkan artefak komputasi yang akan dikonsumsi oleh tahap hilir secara berkesinambungan. Sebagai contoh, fase Data Preparation dan Preprocessing menghasilkan korpus teks bersih yang kemudian diubah menjadi representasi matriks pada tahap Tokenization. Selanjutnya, fase Bayesian Optimization menghasilkan konfigurasi hyperparameter yang paling optimal. Konfigurasi ini secara langsung dikonsumsi oleh fase Fine-Tuning 5-Folds untuk melatih lima arsitektur model secara independen. Fase pelatihan tersebut menghasilkan bobot model beserta metrik validasinya masing-masing, yang kemudian dikonsumsi dan diagregasi oleh arsitektur WPA Ensemble (Weighted Probability Averaging). Terakhir, hasil prediksi probabilitas gabungan tersebut dievaluasi secara ketat pada tahap Evaluation & Benchmarking.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 1. Diagram Alur Penelitian

2.2 Akuisisi dan Pra-pemrosesan Data

Eksperimen komputasional ini memanfaatkan korpus teks gabungan dari kompetisi *Hate Speech and Offensive Content Identification* (HASOC) edisi tahun 2020 dan 2021 untuk menguji ketangguhan model terhadap dinamika pergeseran data secara temporal [11], [12]. Rincian jumlah sampel dari penggabungan *dataset* tersebut disajikan pada Tabel 1, yang menunjukkan adanya *class imbalance* antara sampel kelas mayoritas (teks netral) dan kelas minoritas (ujaran kebencian).

Tabel 1. Distribusi Dataset Gabungan HASOC 2020 + 2021 (Bahasa Inggris)

Task	Label	Keterangan	HASOC 2020	HASOC 2021
1A	HOF	Hate and Offensive	1.856	2.501
	NOT	Non Hate-Offensive	1.852	1.342
1B	HATE	Ujaran Kebencian	158	683
	OFFN	Ofensif	321	622
	PRFN	Profan	1.377	1.196
	NONE	Non Ujaran Kebencian	1.852	1.342
Total	-	-	3.708	3.843

Secara struktural, format atribut dari korpus *raw data* diilustrasikan pada Tabel 2 [12]. Struktur *dataset* ini memuat fitur teks cuitan mentah beserta kolom label hierarkisnya, atribut *task_1* difungsikan sebagai *ground truth* untuk klasifikasi biner, sedangkan *task_2* digunakan untuk klasifikasi *fine-grained* multi-kelas [12]. Untuk memitigasi risiko *data leakage* akibat penggabungan data lintas waktu ini, tahapan deduplikasi diterapkan guna mengeliminasi seluruh sampel teks yang *overlap* atau redundan sebelum dilakukan pemrosesan lebih lanjut [5].

Tabel 2. Data Latih HASOC 2021

No	Teks *tweet*	Task 1A	Task_1B
1	@wealth if you made it through this & were not only able to start making money for yourself but sustain living that way all from home, fuck these companies & corporate pigs. power to the people, always.	HOF	PRFN
2	Technically that's still turning back the clock, dick head https://t.co/jbKaPJmpt1	HOF	OFFN
3	@VMBJP @BJP4Bengal @BJP4India @narendramodi @JPNadda @AmitShah @DilipGhoshBJP @RahulSinhaBJP And you're the govt?!?! Stop thinking about world media, liberal gangs or any optics whatsoever and ACT NOW already. If this is what a person at your level is facing then shudder to think the plight of common people in Bengal. #BengalBurning	NOT	NONE
4	@krtoprak_yigit Soldier of Japan Who has dick head	HOF	OFFN
5	@blueheartedly You'd be better off asking who DOESN'T think he's a sleazy shitbag lmao.	HOF	OFFN

Pada tahap pra-pemrosesan teks, penelitian ini mengadopsi standar pembersihan yang ditetapkan oleh model *cardiffnlp/twitter-roberta-base-offensive* berbasis *framework* TweetEval [13]. Untuk mempertahankan informasi pragmatis dan intensitas emosi pengguna yang penting bagi representasi model, metode pembersihan agresif seperti penghapusan tanda baca atau *stemming* dihindari [14]. Pembersihan teks difokuskan pada penanganan *noise* teknis melalui normalisasi khusus, yaitu dengan mengonversi seluruh *user mentions* menjadi token statis @user dan mereduksi variasi tautan web menjadi token dasar http [13].

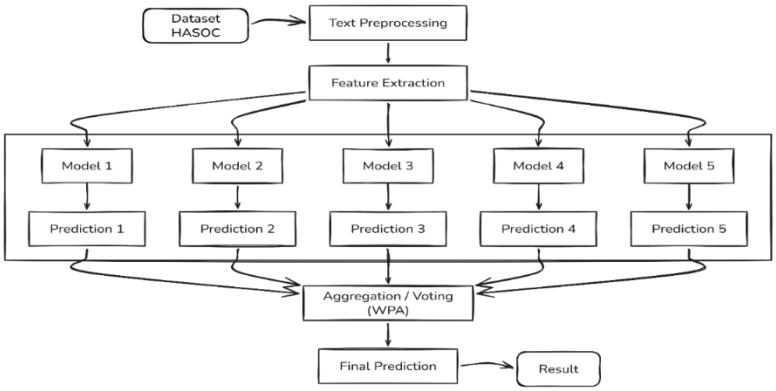
Teks yang telah dinormalisasi kemudian ditransformasikan ke dalam representasi matriks numerik menggunakan tokenisasi *Byte-Pair Encoding* (BPE) tingkat *byte* dengan batas *truncation* maksimal sebesar 128 token [15]. Implementasi tokenisasi ini mampu memecah variasi ejaan kata *slang* internet menjadi unit sub-kata (*subword*) tanpa merusak integritas kontekstual kalimat asli yang diperlukan dalam pemodelan bahasa RoBERTa [15].

2.3 5-Fold Ensemble dengan Weighted Probability Averaging

Gambar 2 mengilustrasikan arsitektur inferensi dari pendekatan pembelajaran *ensemble* yang diusulkan untuk meningkatkan performa generalisasi dibandingkan dengan model tunggal [16]. Pada tahap inferensi, teks masukan (*tweet*)

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

yang telah melalui proses pra-pemrosesan diumpankan secara simultan ke dalam kelima model *fold* secara independen [17]. Setiap model dasar mengekstraksi fitur linguistik secara mandiri dan menghasilkan distribusi probabilitas kelas menggunakan fungsi aktivasi *softmax* [14], [16].



Gambar 2. Arsitektur 5-Fold Ensemble dengan Weighted Probability Averaging (WPA)

Mayoritas sistem *ensemble* konvensional umumnya mengandalkan metode agregasi dasar seperti pemungutan suara mayoritas (*Majority Voting*) atau rata-rata probabilitas seragam (*Unweighted Model Averaging*) [16]. Kendati demikian, kajian literatur menunjukkan bahwa pendekatan rata-rata seragam tersebut rentan menghasilkan prediksi yang kurang optimal karena hasil akhirnya dapat didominasi oleh prediksi dari model yang berkinerja rendah atau mengalami tingkat keyakinan berlebih (*overconfident*) [16]. Kelemahan agregasi konvensional ini menjadi krusial ketika diterapkan pada *dataset* klasifikasi teks yang memiliki tingkat ketidakseimbangan kelas (*class imbalance*) yang tinggi, karena probabilitas keputusan akhir cenderung memihak pada kelas mayoritas [1].

Sebagai solusi untuk memitigasi bias tersebut, vektor probabilitas dari masing-masing model dikombinasikan secara dinamis menggunakan teknik *Weighted Probability Averaging* (WPA) [1], [18]. Pendekatan ini bekerja dengan memberikan *weight* yang dihitung secara proporsional berdasarkan metrik performa evaluasi *Macro F1-Score* yang diraih oleh masing-masing model dasar pada fase validasi [1], [18]. Strategi pembobotan probabilistik ini dirancang untuk memastikan bahwa model dengan kemampuan prediksi yang lebih superior secara otomatis memberikan pengaruh yang lebih besar pada keputusan klasifikasi akhir [1], [18].

Pada korpus deteksi ujaran kebencian yang memiliki ketimpangan kelas yang tinggi, kalibrasi bobot dinamis ini secara signifikan dapat meningkatkan sensitivitas sistem dalam mendeteksi kelas minoritas dengan cara meminimalkan dampak klasifikasi dari model yang memiliki tingkat keyakinan rendah [18], [19]. Sebagai langkah pemrosesan tambahan, tahapan optimasi ambang batas keputusan (*threshold tuning*) dapat diterapkan pada hasil agregasi WPA untuk memaksimalkan performa *Macro F1-Score* akhir sebelum penentuan label kelas secara definitif [16]. Secara matematis, proses normalisasi bobot berdasarkan performa validasi model direpresentasikan melalui Persamaan (1), Selanjutnya, agregasi probabilitas dinamis dari seluruh model anggota dalam metode WPA diformulasikan melalui Persamaan (2) di mana $P_{ensemble}(c|x)$ merupakan probabilitas prediksi akhir arsitektur ensemble untuk kelas target c dengan kondisi input x [18]. Variabel K menunjukkan jumlah total model dasar, sedangkan $P_k(c|x)$ adalah probabilitas keluaran yang diprediksi oleh model ke-k. Bobot spesifik model W_k dinormalisasi berdasarkan variabel $F1_k$, yaitu nilai Macro F1-Score validasi yang dicapai secara empiris oleh model ke-k [18].

$$P_{\text{ensemble}}(c \mid x) = \sum_{k=1}^{K} W_k \cdot P_k(c \mid x) \quad (1)$$

$$W_k = \frac{F1_k}{\sum_{j=1}^{K} F1_j} \quad (2)$$

2.4 Skenario Eksperimen dan Optimasi Bayesian

Evaluasi komprehensif terhadap arsitektur klasifikasi yang diusulkan dieksekusi melalui empat skenario eksperimen progresif yang secara ketat mengadopsi paradigma ablation study, sebagaimana dirincikan pada Tabel 3 [1]. Pendekatan eksperimental ini dikonstruksi dengan cara mengintegrasikan tepat satu intervensi metodologis secara bertahap pada setiap skenarionya guna mengisolasi variabel komputasional yang diuji [8]. Desain pengujian metodologis tersebut secara analitis memungkinkan observasi atribusi kausal yang presisi terkait bagaimana setiap penambahan komponen algoritma memengaruhi fluktuasi performa prediksi model secara keseluruhan [20].

Tabel 3. Konfigurasi Skenario Eksperimen

Skenario	Dataset	Hyperparameter	Ensemble
S1 (Baseline)	HASOC 2021 saja	Default (LR=5e-5)	Model tunggal
S2	2020 + 2021 digabung	Default (LR=5e-5)	Model tunggal
S3	2020 + 2021 digabung	Bayesian (Optuna)	Model tunggal
S4 (Final)	2020 + 2021 digabung	Bayesian (Optuna)	5-Fold WPA Ensemble

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

Untuk memaksimalkan performa konvergensi pada Skenario 3 (S3) dan Skenario 4 (S4), penelitian ini mengimplementasikan algoritma *Bayesian Optimization* melalui *framework* Optuna menggunakan pendekatan *Tree-structured Parzen Estimator* (TPE) [6]. Berbeda dengan *Grid Search* konvensional yang mengevaluasi kombinasi parameter secara menyeluruh, estimator probabilistik ini mengonstruksi model *surrogate* untuk mengeksplorasi ruang *hyperparameter* secara efisien [6]. Optimasi ini difokuskan pada pencarian *learning rate* logaritmik dan *batch size* guna memitigasi instabilitas *fine-tuning* serta anomali *vanishing gradient* yang kerap memengaruhi keberhasilan pelatihan pada arsitektur Transformer [6].

Tabel 4. Ruang Pencarian Optimasi Bayesian

Parameter	Rentang Pencarian	Skala
Learning Rate	1×10^{-5} - 5×10^{-5}	Logaritmik
Warmup Ratio	0,05 - 0,15	Linear
Weight Decay	1×10^{-3} - 1×10^{-1}	Logaritmik

Berdasarkan Tabel 4, batasan *learning rate* ditetapkan pada nilai yang relatif rendah, yaitu maksimal 5×10^{-5}. Pembatasan ini dilakukan karena penggunaan nilai yang lebih tinggi terbukti secara empiris dapat menyebabkan kegagalan konvergensi pada arsitektur Transformer [7], [21]. Fungsi objektif yang digunakan dalam algoritma pencarian ini adalah maksimasi nilai *Macro F1-Score* pada data validasi. Melalui penerapan *Bayesian Optimization*, arsitektur model dapat menentukan konfigurasi optimal secara adaptif dengan jumlah iterasi (*trials*) yang lebih sedikit dibandingkan dengan metode konvensional. Pendekatan ini secara signifikan meningkatkan efisiensi waktu komputasi sekaligus menjaga stabilitas pelatihan model sebelum memasuki fase *5-Fold Ensemble*.

2.5 Evaluasi *Blind Test* dan *Benchmarking*

Tahap akhir dari metodologi ini adalah evaluasi performa arsitektur ensemble menggunakan skema blind test pada dataset pengujian yang tidak terlihat selama fase pelatihan [14]. Evaluasi independen ini sangat krusial untuk memastikan bahwa model mampu melakukan generalisasi prediksi secara objektif di dunia nyata dan tidak sekadar menghafal pola data latih [14]. Dalam konteks deteksi ujaran kebencian yang sering mengalami masalah ketidakseimbangan kelas secara ekstrem, penggunaan metrik akurasi dasar dapat memberikan ilusi performa yang menyesatkan karena hasil klasifikasi akan cenderung memihak pada prediksi kelas mayoritas [1]. Oleh karena itu, kinerja akhir model pada penelitian ini dievaluasi secara spesifik menggunakan metrik Macro F1-Score [22]. Metrik ini ditetapkan sebagai tolok ukur penentu karena perhitungannya memberikan bobot yang seimbang untuk setiap kelas secara independen, sehingga kemampuan model dalam mendeteksi kelas minoritas diperlakukan setara dengan kelas mayoritas [22]. Kemampuan prediksi akhir dari arsitektur ensemble ini kemudian akan diuji daya saingnya melalui proses benchmarking terhadap papan peringkat resmi dari kompetisi global HASOC [12].

Secara matematis, perumusan untuk menghitung evaluasi kinerja klasifikasi model tersebut disajikan pada Persamaan (3) [22]. Berdasarkan persamaan tersebut, variabel *C* merepresentasikan jumlah total kelas klasifikasi yang terdapat pada dataset [22]. Sementara itu, variabel F1-i merepresentasikan nilai skor evaluasi F1-Score individual yang berhasil dicapai oleh model untuk kelas ke-i [22]. Dengan membagi total penjumlahan F1-Score dari seluruh kelas dengan jumlah kelas tersebut, sistem dapat menghasilkan satu nilai rata-rata makro yang merepresentasikan ketangguhan model secara keseluruhan tanpa terpengaruh oleh dominasi frekuensi kelas tertentu [22].

$$Macro\ F1\text{-}Score = \frac{1}{|C|}\sum_{c\in C} F1_c \quad (3)$$

3. HASIL DAN PEMBAHASAN

3.1 Proses dan Hasil Optimasi Parameter Berbasis *Bayesian*

Tantangan utama dalam melakukan proses *fine-tuning* pada arsitektur Transformer menggunakan korpus *dataset* turunan (*downstream*) yang terbatas adalah tingginya kerentanan model terhadap instabilitas optimasi dan variansi performa yang ekstrem [7], [8]. Untuk mengatasi kendala komputasi tersebut, penelitian ini mengimplementasikan algoritma *Bayesian Optimization* melalui pustaka Optuna guna memandu pencarian *hyperparameter* secara adaptif dan terarah [10].

Berbeda dengan metode *Grid Search* konvensional yang mengevaluasi setiap kombinasi parameter secara menyeluruh sehingga membutuhkan sumber daya komputasi yang besar, estimator Bayesian bekerja dengan mengonstruksi model probabilitas *surrogate* yang diperbarui secara berulang berdasarkan riwayat evaluasi eksperimen sebelumnya [6]. Secara spesifik, pendekatan *Tree-structured Parzen Estimator* (TPE) yang terintegrasi di dalam Optuna mampu mengeliminasi konfigurasi berkinerja rendah sejak dini dan memfokuskan iterasi pencarian pada ruang parameter yang berpotensi menghasilkan tingkat konvergensi optimal. Hasil ekstraksi *hyperparameter* terbaik dari proses optimasi probabilistik pada Skenario 4 untuk tugas *Subtask* 1A dan *Subtask* 1B disajikan secara rinci pada Tabel 5 [6].

Tabel 5. Hasil Optimasi Bayesian - Hyperparameter Optimal (15 Percobaan)

Hyperparameter	Subtask 1A (Biner)	Subtask 1B (Multi-kelas)
Learning Rate	$2{,}23 \times 10^{-5}$	$4{,}20 \times 10^{-5}$

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

Hyperparameter	Subtask 1A (Biner)	Subtask 1B (Multi-kelas)
Warmup Ratio	0,0549	0,1144
Weight Decay	0,029883	0,001652

Berdasarkan data pada Tabel 5, proses optimasi probabilistik secara spesifik menghasilkan konfigurasi *learning rate* yang konservatif, yakni $2{,}23 \times 10^{-5}$ untuk *Subtask* 1A dan $4{,}20 \times 10^{-5}$ untuk *Subtask* 1B [7]. Penetapan nilai *learning rate* yang relatif kecil ini merupakan strategi mitigasi karena lapisan representasi bawah dari arsitektur Transformer rentan mengalami fenomena *vanishing gradient* apabila dieksekusi dengan pembaruan bobot yang terlalu agresif di awal pelatihan [7]. *Learning rate* yang melewati batas toleransi dapat merusak representasi linguistik yang telah dipelajari oleh model selama fase *pre-training* pada miliaran korpus teks sebelumnya. Anomali komputasional ini dikenal secara luas dalam bidang *Natural Language Processing* (NLP) sebagai *catastrophic forgetting* [7], [23].

Lebih lanjut, algoritma Bayesian juga mengonfigurasi parameter *warmup ratio* sebesar 0,0549 dan 0,1144, yang secara operasional berfungsi untuk menekan ukuran langkah gradien mendekati nol pada iterasi awal pelatihan [7]. Penerapan *warmup* yang dipadukan dengan kalibrasi *weight decay* ini memberikan ruang bagi *optimizer* untuk menstabilkan pergerakan arah vektor bobot sebelum mempercepat laju pembaruan parameter, sehingga secara empiris mampu menghindarkan model dari hambatan konvergensi lokal [7]. Melalui perpaduan mekanisme tersebut, intervensi *Bayesian Optimization* ini tidak hanya meningkatkan efisiensi waktu pencarian parameter, tetapi juga menjamin stabilitas struktural model RoBERTa dalam mempertahankan akurasi generalisasi terhadap dinamika data ujaran kebencian di dunia nyata [8], [10].

3.2 Analisis Stabilitas Pelatihan 5-Fold Cross-Validation

Tahap komputasi lanjutan setelah penentuan *hyperparameter* berfokus pada pelatihan arsitektur klasifikasi menggunakan skema *Stratified 5-Fold Cross-Validation* untuk memitigasi variansi performa yang lazim terjadi pada pemrosesan model bahasa berbasis Transformer [24]. Secara teoretis, proses *fine-tuning* pada arsitektur Transformer sangat sensitif terhadap inisialisasi acak (*random seed*) dan urutan distribusi data latih. Perbedaan inisialisasi pada *dataset* berukuran kecil dapat memicu kegagalan optimasi secara tiba-tiba [24].

Pendekatan pemisahan data berbasis stratifikasi ini diterapkan untuk memastikan bahwa setiap partisi pelatihan maupun validasi mempertahankan proporsi distribusi kelas minoritas yang identik dengan populasi aslinya. Hal ini dilakukan untuk mencegah bias representasi yang dapat menurunkan kemampuan generalisasi model [24]. Hasil evaluasi metrik validasi dari kelima model *fold* yang dilatih secara independen untuk tugas *Subtask* 1A dan *Subtask* 1B dirangkum pada Tabel 6.

Tabel 6. Macro F1-Score Validasi dan Bobot WPA per Fold

Model Fold	Macro F1 Val. (1A)	Bobot WPA (1A)	Macro F1 Val. (1B)	Bobot WPA (1B)
Fold 1	0,8646	0,2002	0,6654	0,2005
Fold 2	0,8638	0,2000	0,6835	0,2059
Fold 3	0,8726	0,2020	0,6352	0,1914
Fold 4	0,8566	0,1982	0,6731	0,2028
Fold 5	0,8592	0,1990	0,6620	0,1995
Rata-rata +/- SD	0,8634 ± 0, 0060	1,0000	0,6638 ± 0,0161	1,0000

Berdasarkan data pada Tabel 6, hasil evaluasi empiris menunjukkan bahwa perpaduan antara manajemen parameter *seed* statis dan pemisahan data berimbang efektif dalam menjaga stabilitas konvergensi model pada kedua tugas klasifikasi tersebut [24]. Pada *Subtask* 1A, model menghasilkan nilai standar deviasi yang kecil, yaitu ± 0,0060. Hal ini mengindikasikan bahwa seluruh model *fold* mampu menggeneralisasi fitur ujaran kebencian secara konsisten tanpa mengalami degradasi performa pada *fold* tertentu [24].

Konsistensi stabilitas tersebut juga terlihat pada *Subtask* 1B yang memiliki kompleksitas ambiguitas semantik lebih tinggi, dengan standar deviasi sebesar ± 0,0161 meskipun dihadapkan pada tumpang tindih leksikal antarkategori kelas yang berbeda [24]. Selanjutnya, skor *Macro F1-Score* validasi dari kelima model independen ini ditransformasikan menjadi bobot probabilitas dinamis dalam algoritma *Weighted Probability Averaging* (WPA) [24].

Melalui mekanisme pembobotan ini, model yang menunjukkan kinerja lebih tinggi seperti *Fold* 3 pada *Subtask* 1A (bobot 0,2020) dan *Fold* 2 pada *Subtask* 1B (bobot 0,2059) diberikan kontribusi yang lebih besar dalam penentuan keputusan keputusan akhir. Sebaliknya, pengaruh dari model dengan kinerja yang lebih rendah dibatasi (seperti *Fold* 3 pada *Subtask* 1B dengan bobot 0,1914). Melalui metode ini, arsitektur *ensemble* WPA dapat meminimalkan dampak prediksi yang keliru untuk menjamin ketangguhan klasifikasi pada lingkungan *noisy data* [24].

3.3 Hasil Evaluasi Metrik Prediksi Klasifikasi

Evaluasi performa arsitektur ensemble WPA menggunakan data blind test independen menunjukkan kemampuan generalisasi yang sangat kuat pada klasifikasi biner Subtask 1A, sebagaimana dirangkum secara kuantitatif pada Tabel 7 [12]. Sistem berhasil mencatatkan tingkat Recall yang sangat tinggi (0,9261) pada kelas netral (NOT), yang membuktikan efektivitas komputasional model dalam menekan rasio False Positive guna mencegah pemblokiran berlebihan pada percakapan normal [2]. Meskipun tingkat kepekaan pada kelas ujaran kebencian (HOF) terdeteksi lebih rendah akibat

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

kerumitan bahasa kiasan dan sarkasme, pencapaian Macro F1-Score agregat sebesar 0,8099 menegaskan bahwa penggabungan probabilitas berbobot sukses mempertahankan ketangguhan prediksi model [5].

Tabel 7. Laporan Klasifikasi Subtask 1A (Ensemble WPA, Blind Test)

Kelas	Presisi	Recall	F1-Score	Support
NOT	0,8229	0,9261	0,8715	798
HOF	0,8460	0,6798	0,7483	483
Akurasi	-	-	0,8298	1.281
Macro Avg	0,8344	0,8099	0,8099	1.281

Kompleksitas prediksi komputasional meningkat secara drastis pada klasifikasi fine-grained untuk Subtask 1B, dengan rincian metrik performa yang disajikan pada Tabel 8 [5]. Arsitektur ini paling sukses mengidentifikasi kelas profan (PRFN) dengan F1-Score tertinggi (0,7750) berkat keberadaan pola leksikal umpatan kasar yang relatif konsisten dan mudah ditangkap oleh ekstraksi fitur [9]. Sebaliknya, tingginya tingkat tumpang tindih fitur linguistik antara serangan personal pada kelas ofensif (OFFN) dan serangan target demografis pada kelas ujaran kebencian (HATE) memicu kesalahan klasifikasi yang menekan performa metrik pada kelas OFFN [5]. Walaupun terhambat oleh tingkat ambiguitas semantik tersebut, perolehan Macro F1-Score agregat sebesar 0,6470 membuktikan secara empiris bahwa metode WPA efektif menyeimbangkan bias antar-kelas pada lingkungan dataset yang distribusinya sangat timpang [16].

Tabel 8. Laporan Klasifikasi Subtask 1B (Ensemble WPA, Blind Test)

Kelas	**Presisi**	**Recall**	**F1-Score**	**Support**
HATE	0,5662	0,6250	0,5920	224
OFFN	0,4917	0,4564	0,4734	195
PRFN	0,7000	0,8681	0,7750	379
NONE	0,8478	0,6687	0,7477	483
Akurasi	-	-	0,6877	1.281
Macro Avg	0,6504	0,6546	0,6470	1.281

3.4 Hasil Studi Ablasi Terhadap Pergeseran Distribusi Data

Evaluasi dampak kausal dari setiap intervensi algoritmik terhadap fenomena distributional shift dianalisis secara terukur melalui empat skenario studi ablasi progresif, sebagaimana dirangkum pada Tabel 9 [1]. Transisi evaluasi dari Skenario 1 (S1) menuju Skenario 2 (S2) memperlihatkan degradasi performa klasifikasi yang tajam dari 81,73% menjadi 80,03% akibat penggabungan korpus lintas tahun [5]. Penurunan metrik tersebut mengonfirmasi bahwa arsitektur model Transformer tunggal sangat rapuh ketika dihadapkan pada evolusi temporal, karena model cenderung terjebak dalam bias hafalan (in-distribution overfitting) terhadap leksikon atau tagar spesifik pada satu periode waktu tertentu [7].

Tabel 9. Perbandingan Skenario Progresif - Macro F1-Score pada Blind Test

Skenario	**Intervensi**	**F1 (1A)**	**F1 (1B)**
S1	Model tunggal + HASOC 2021 saja (parameter default)	0,8173	0,6574
S2	Model tunggal + Gabungan 2020+2021 (parameter default)	0,8003	0,6554
S3	Model tunggal + Gabungan + Bayesian Optimization	0,8072	0,6404
S4	5-Fold Ensemble WPA + Gabungan + Bayesian Optim.	0,8099	0,6470

Untuk mengatasi degradasi tersebut, penerapan algoritma Bayesian Optimization pada Skenario 3 (S3) terbukti mampu memulihkan performa menjadi 80,72% berkat kemampuannya dalam menstabilkan aliran gradien selama proses fine-tuning pada lingkungan data yang bising [7]. Selanjutnya, aktivasi arsitektur ensemble dengan mekanisme agregasi Weighted Probability Averaging (WPA) pada Skenario 4 (S4) berhasil mengatrol dan menstabilkan performa klasifikasi secara komprehensif hingga mencapai titik optimal sebesar 80,99% [1]. Secara empiris, meskipun skor akhir S4 tampak sedikit di bawah baseline S1 yang terindikasi mengalami cacat overfitting pada memori spasialnya, arsitektur integratif S4 merepresentasikan sistem yang secara fundamental jauh lebih tangguh dalam mempertahankan akurasi generalisasi pada dinamika wacana ujaran kebencian di dunia nyata [16].

3.5 Benchmarking Performa Klasifikasi di Tingkat Global

Tahap evaluasi komputasional akhir dilakukan melalui proses uji banding (*benchmarking*) terhadap papan peringkat (*leaderboard*) resmi dari kompetisi internasional *Hate Speech and Offensive Content Identification* (HASOC) edisi 2021 untuk menguji daya saing arsitektur yang diusulkan [12]. Kompetisi global tersebut melibatkan partisipasi dari 65 tim penelitian dari berbagai negara, sehingga menjadi tolok ukur (*gold standard*) yang relevan untuk memvalidasi ketangguhan inovasi algoritma deteksi ujaran kebencian [12]. Hasil pembandingan performa untuk tugas klasifikasi biner pada *Subtask* 1A disajikan secara rinci pada Tabel 10 [12]. Berdasarkan data pada tabel tersebut, capaian *Macro F1-Score* sebesar 0,8099 dari sistem *ensemble* WPA yang diusulkan mampu melampaui metrik milik tim HNLP yang secara resmi menduduki peringkat keempat global [12].

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

Tabel 10. Benchmarking Arsitektur yang Diusulkan dengan Leaderboard Resmi HASOC 2021 Subtask 1A

Peringkat	Tim / Arsitektur	F1-Score
1	NLP-CIC	0,8305
2	HUNLP	0,8215
3	neuro-utmn-thales	0,8199
-*	Usulan: Bayesian + 5-Fold WPA Ensemble	0,8099
4	HNLP	0,8089
5	Chandigarh_Concordia	0,8040
6	KuiYongyi	0,8030

Tren performa tingkat elit ini juga terekam secara konsisten pada tugas klasifikasi multi-kelas untuk Subtask 1B yang memiliki kompleksitas semantik jauh lebih tinggi, sebagaimana dirangkum secara kuantitatif pada Tabel 11 [5], [12]. Eksekusi arsitektur ensemble pada sub-tugas tersebut sukses membukukan metrik sebesar 0,6470, yang secara efektif memposisikan sistem ini di atas tim Super Mario yang menempati peringkat keempat resmi [12]. Pencapaian yang berhasil mengamankan posisi top-4 global secara tidak resmi di kedua sub-tugas kompetitif ini menegaskan bahwa kerentanan prediksi pada model bahasa tunggal dapat diatasi secara komprehensif melalui strategi agregasi ensemble probabilistik [9].

Tabel 11. Benchmarking Arsitektur yang Diusulkan dengan Leaderboard Resmi HASOC 2021 Subtask 1B

Peringkat	Tim / Arsitektur	F1 (1B)
1	NLP-CIC	0,6657
2	neuro-utmn-thales	0,6577
3	HASOC21rub	0,6482
-*	Usulan: Bayesian + 5-Fold WPA Ensemble	0,6470
4	Super Mario	0,6447
5	UINSUSKA	0,6417
6	HNLP	0,6396

3.6 Pembahasan dan Perbandingan Performa

Hasil eksperimen secara keseluruhan mengonfirmasi bahwa arsitektur *5-Fold Ensemble* dengan *Weighted Probability Averaging* (WPA) yang diusulkan mampu mengatasi keterbatasan pendekatan model tunggal konvensional. Jika dibandingkan dengan penelitian terdahulu oleh Mozafari dkk. (2020) yang hanya menerapkan *fine-tuning* pada model bahasa BERT tunggal untuk mendeteksi ujaran kebencian [25], sistem yang dikembangkan dalam penelitian ini menunjukkan ketangguhan yang lebih superior terhadap masalah ketidakseimbangan kelas ekstrem.

Lebih lanjut, dalam sebuah tinjauan komprehensif terkait teknik *ensemble deep learning*, Ganaie dkk. (2022) menyatakan bahwa mayoritas metode *ensemble* konvensional masih sangat bergantung pada strategi agregasi dasar seperti *Majority Voting* [16]. Arsitektur usulan dalam penelitian ini membuktikan bahwa metode *Majority Voting* tersebut rentan terhadap bias kelas mayoritas. Kendala tersebut secara empiris dapat diatasi melalui kalkulasi WPA yang meminimalkan bias dengan memberikan bobot dinamis yang lebih besar kepada model anggota berdasarkan metrik *F1-Score* validasi spesifik masing-masing *fold* [16].

Dalam konteks stabilitas komputasional, penelitian ini mengatasi masalah instabilitas *fine-tuning* pada model Transformer yang sebelumnya disoroti oleh Mosbach dkk. (2021) dan Dodge dkk. (2020) [7], [8]. Ketika penelitian terdahulu umumnya mengandalkan metode pencarian *Grid Search* atau *Random Search* yang membutuhkan biaya komputasi tinggi dan berisiko melewatkan titik optimal [8], penelitian ini sejalan dengan temuan Ridwan dan Utami (2024) yang membuktikan bahwa *Bayesian Optimization* melalui *framework* Optuna mampu menemukan ruang parameter secara lebih efisien dan presisi [21].

Kebaruan dan letak perbedaan penelitian ini terletak pada keberhasilan penerapan Optuna secara khusus untuk memitigasi anomali *vanishing gradient* pada arsitektur *Twitter-RoBERTa*. Hasil optimasi mengonfirmasi bahwa penggunaan nilai *learning rate* yang rendah, yaitu berkisar antara 2e-05 hingga 4e-05, merupakan faktor krusial untuk mempertahankan integritas representasi linguistik selama proses *fine-tuning* pada teks ujaran kebencian [7].

Dari sisi ketangguhan generalisasi dan evaluasi komparatif global, performa arsitektur usulan dianalisis secara kontras terhadap sistem-sistem pada kompetisi HASOC 2021 [12]. Berdasarkan tinjauan literatur dari penyelenggara resmi, tim-tim yang mendominasi peringkat puncak dalam deteksi ujaran kebencian umumnya bergantung pada injeksi korpus data eksternal berskala besar (seperti *HatebaseTwitter*) serta menggunakan kombinasi arsitektur multibahasa raksasa yang membutuhkan kapasitas memori yang besar [12]. Sebaliknya, arsitektur usulan dalam penelitian ini mampu mencapai posisi kompetitif elit dengan menempati posisi *top-4* global secara tidak resmi secara efisien melalui optimalisasi fitur bawaan *Twitter-RoBERTa* monolingual tanpa mengandalkan tambahan data dari luar [12].

Perbandingan ini membuktikan secara ilmiah bahwa strategi pemisahan distribusi data melalui *Stratified 5-Fold*, yang dipadukan dengan agregasi probabilistik WPA dan optimasi *hyperparameter* berbasis Bayesian, mampu menghasilkan model yang lebih efisien, terukur, dan tangguh secara komputasional dibandingkan pendekatan konvensional yang menitikberatkan pada peningkatan skala arsitektur jaringan secara masif.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 199–208
DOI : 10.47065/bulletinds.v5i3.10145
https://ejurnal.seminar-id.com/index.php/bulletinds

4. KESIMPULAN

Berdasarkan keseluruhan rangkaian eksperimen dan evaluasi yang telah dilaksanakan, penelitian ini berhasil merancang, mengimplementasikan, dan memvalidasi arsitektur *pipeline Natural Language Processing* (NLP) *end-to-end* yang *robust* untuk mendeteksi ujaran kebencian di media sosial. Secara empiris, penggabungan korpus lintas tahun dari *dataset* HASOC edisi 2020 dan 2021 terbukti menghasilkan *noise* distribusional dan memicu fenomena *distributional shift* yang secara signifikan menurunkan kemampuan prediksi arsitektur model Transformer tunggal konvensional. Untuk mengatasi kendala optimasi tersebut, penerapan *Bayesian Optimization* melalui *framework* Optuna mampu mengarahkan pencarian ruang *hyperparameter* secara adaptif. Langkah ini menghasilkan konfigurasi parameter yang menstabilkan proses *fine-tuning* pada model basis *Twitter-RoBERTa* sekaligus memitigasi terjadinya anomali *vanishing gradient*. Lebih lanjut, penerapan metode *Stratified 5-Fold Cross-Validation* yang diintegrasikan dengan teknik agregasi *Weighted Probability Averaging* (WPA) menjadi intervensi penting dalam tahapan ini. Berbeda dengan mekanisme agregasi konvensional, kalibrasi probabilitas dinamis pada WPA secara efektif meminimalkan bias terhadap kelas mayoritas dengan memberikan bobot keputusan prediktif yang lebih besar secara proporsional kepada model *fold* yang memiliki performa validasi tertinggi. Sinergi dari seluruh tahapan metodologis ini memberikan hasil yang konkret dan kompetitif, di mana sistem *ensemble* yang diusulkan berhasil mencapai nilai *Macro F1-Score* sebesar 80,99% untuk tugas klasifikasi biner pada *Subtask* 1A, serta 64,70% untuk tantangan klasifikasi *fine-grained* multi-kelas pada *Subtask* 1B. Pencapaian komprehensif tersebut tidak hanya mengungguli performa model *baseline*, melainkan juga menempatkan arsitektur ini pada posisi *top-4* global saat diuji banding (*benchmarking*) dengan *leaderboard* resmi kompetisi internasional HASOC 2021. Sebagai kesimpulan, integrasi antara optimasi probabilistik dan agregasi berbobot ini menawarkan fondasi arsitektur moderasi konten digital yang stabil dan dapat direproduksi (*reproducible*) untuk implementasi pada kondisi riil. Penelitian selanjutnya disarankan untuk mengarah pada pengembangan pemrosesan data multimodal serta adaptasi komputasional terhadap bahasa-bahasa daerah yang memiliki keterbatasan sumber daya linguistik.

REFERENCES

[1] S. Yoo, E. Jeon, J. Hyeon, and J. Cho, "Adaptive ensemble techniques leveraging BERT based models for multilingual hate speech detection in Korean and english," *Sci. Rep.*, vol. 15, no. 1, p. 19844, Jun. 2025, doi: 10.1038/s41598-025-88960-y.

[2] D. Hartmann, A. Oueslati, D. Staufer, L. Pohlmann, S. Munzert, and H. Heuer, "Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2025, pp. 1–26. doi: 10.1145/3706598.3713998.

[3] A. Kumar, P. K. Roy, and S. Saumya, "An Ensemble Approach for Hate and Offensive Language Identification in English and Indo-Aryan Languages," *Sci. Rep* 2021. Accessed: Dec. 07, 2025. [Online]. Available: https://ceur-ws.org/Vol-3159/T1-43.pdf

[4] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," *Sci. Rep* pp. 9–14, Oct. 2020, Accessed: Dec. 07, 2025. [Online]. Available: http://arxiv.org/abs/2005.10200

[5] D. Antypas and J. Camacho-Collados, "Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation," *Sci. Rep* Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.01680

[6] X. Liu and C. Wang, "An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 2286–2300. doi: 10.18653/v1/2021.acl-long.178.

[7] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2006.04884

[8] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.06305

[9] I. E. Kucukkaya and C. Toraman, "Constructing ensembles for hate speech detection," *Natural Language Processing*, vol. 31, no. 3, pp. 745–770, May 2025, doi: 10.1017/nlp.2024.44.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.10902

[11] T. Mandla *et al.*, "Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.05927

[12] T. Mandl *et al.*, "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.09301

[13] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of the Association for Computational Linguistics:*

EMNLP 2020, Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2020, pp. 1644–1650. doi: 10.18653/v1/2020.findings-emnlp.148.

[14] M. S. Jinan, M. R. Handayani, M. A. Ulinuha, and K. Umam, "DETEKSI CYBERBULLYING MULTIKELAS BERKINERJA TINGGI: ENSEMBLE ROBERTA-LARGE DENGAN PRESISI CAMPURAN," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 3, pp. 2666–2678, Sep. 2025, doi: 10.29100/jipi.v10i3.8056.

[15] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[16] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, Oct. 2022, doi: 10.1016/j.engappai.2022.105151.

[17] A. C. Mazari, N. Boudoukhani, and A. Djeffal, "BERT-based ensemble learning for multi-aspect hate speech detection," *Cluster Comput.*, vol. 27, no. 1, pp. 325–339, Feb. 2024, doi: 10.1007/s10586-022-03956-x.

[18] N. P. Shetty, Y. Singh, V. Hegde, D. Cenitta, and D. K, "Exploring emotional patterns in social media through NLP models to unravel mental health insights," *Healthc. Technol. Lett.*, vol. 12, no. 1, Jan. 2025, doi: 10.1049/htl2.12096.

[19] 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, p. e3214, Oct. 2025, doi: 10.7717/peerj-cs.3214.

[20] Z. Zhang, J. Chen, and D. Yang, "Mitigating Biases in Hate Speech Detection from A Causal Perspective," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 6610–6625. doi: 10.18653/v1/2023.findings-emnlp.440.

[21] E. M. Pusung and I. N. Dewi, "Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi Emosi berbasis Teks," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 10, no. 3, pp. 240–248, Feb. 2025, doi: 10.25077/TEKNOSI.v10i3.2024.240-248.

[22] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2202.09517

[23] C. Lee, K. Cho, and W. Kang, "Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models," Jan. 2020, [Online]. Available: http://arxiv.org/abs/1909.11299

[24] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, Accessed: Mar. 11, 2026. [Online]. Available: http://arxiv.org/abs/1705.07874

[25] B. Aklouche, Y. Bazine, and Z. Ghalia-Bououchma, "Offensive Language and Hate Speech Detection Using Transformers and Ensemble Learning Approaches," *Computación y Sistemas*, vol. 28, no. 3, pp. 1031–1039, Sep. 2024, doi: 10.13053/cys-28-3-4724.