BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Optimasi Klasifikasi Hate Speech dan Offensive Language melalui Frozen RoBERTa Feature Extraction dan Random Forest

Marsha Cahyani Dwisyakilla, Surya Agustian*, Novriyanto, Muhammad Affandes

Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12250123947@students.uin-suska.ac.id, [2,*]surya.agustian@uin-suska.ac.id, [3]novriyanto@uin-suska.ac.id, [4]affandes@uin-suska.ac.id
Email Penulis Korespondensi: surya.agustian@uin-suska.ac.id

Abstrak−Deteksi *hate speech* dan *offensive content* pada media sosial merupakan tantangan penting dalam pemrosesan bahasa alami karena karakteristik teks Twitter yang singkat, informal, serta mengandung berbagai elemen seperti *mention*, URL, *hashtag*, dan ekspresi emosional yang menyulitkan proses klasifikasi. Pendekatan *fine-tuning end-to-end* berbasis Transformer umumnya membutuhkan sumber daya komputasi yang lebih besar, sehingga penelitian ini mengeksplorasi pendekatan yang lebih efisien melalui pemanfaatan RoBERTa sebagai *frozen feature extractor* yang dikombinasikan dengan Random Forest sebagai klasifikator. Pendekatan ini memungkinkan pemanfaatan representasi kontekstual dari model Transformer tanpa proses pelatihan ulang penuh. Penelitian menggunakan dataset HASOC 2021 English Track yang terdiri atas dua tugas klasifikasi, yaitu Task A untuk klasifikasi biner HOF dan NOT serta Task B untuk klasifikasi multi-kelas HATE, OFFN, PRFN, dan NONE. Optimasi *pipeline* dilakukan melalui penambahan *handcrafted features*, *oversampling*, *hyperparameter tuning* Random Forest, serta *threshold tuning* pada skenario tertentu. Evaluasi dilakukan menggunakan *accuracy*, *precision*, *recall*, dan F1-macro sebagai metrik utama karena distribusi kelas yang tidak seimbang.Hasil terbaik menunjukkan F1-macro sebesar 0,80 pada Task A dan 0,64 pada Task B. Hasil ini menunjukkan bahwa pendekatan *frozen* RoBERTa dan Random Forest mampu memberikan performa yang baik pada klasifikasi biner *hate speech* dan *offensive content*. Namun, performa pada Task B masih menunjukkan keterbatasan dalam membedakan kelas-kelas yang memiliki karakteristik linguistik berdekatan, seperti HATE, OFFN, dan PRFN, sehingga klasifikasi multi-kelas yang lebih rinci masih menjadi tantangan. Secara umum, hasil penelitian mengindikasikan bahwa *frozen feature extraction* berbasis RoBERTa merupakan alternatif yang efisien secara komputasi untuk deteksi *hate speech* pada data Twitter berbahasa Inggris, meskipun pengembangan lebih lanjut masih diperlukan untuk meningkatkan performa pada klasifikasi multi-kelas.

Kata Kunci: HASOC 2021; Ujaran Kebencian; Konten Ofensif; RoBERTa; Random Forest; Fitur Handcrafted

Abstract−Hate speech and offensive content detection on social media remains a significant challenge in Natural Language Processing (NLP) due to the characteristics of Twitter data, which are typically short, informal, and contain various elements such as mentions, URLs, hashtags, and emotional expressions that complicate the classification process. End-to-end Transformer fine-tuning approaches generally require substantial computational resources; therefore, this study explores a more computationally efficient approach by utilizing RoBERTa as a frozen feature extractor combined with Random Forest as the classifier. This approach enables the exploitation of contextual representations generated by Transformer models without requiring full model retraining.The study employs the HASOC 2021 English Track dataset, which consists of two classification tasks: Task A for binary classification (HOF and NOT) and Task B for multi-class classification (HATE, OFFN, PRFN, and NONE). The classification pipeline is optimized through the incorporation of handcrafted features, oversampling, Random Forest hyperparameter tuning, and threshold tuning in specific scenarios. Model performance is evaluated using accuracy, precision, recall, and F1-macro, with F1-macro serving as the primary metric due to class imbalance. The best-performing model achieved an F1-macro score of 0.80 on Task A and 0.64 on Task B. These results indicate that the combination of frozen RoBERTa representations and Random Forest provides strong performance for binary hate speech and offensive content classification. However, the performance on Task B highlights the difficulty of distinguishing linguistically similar categories, such as HATE, OFFN, and PRFN, suggesting that fine-grained multi-class classification remains a challenging task. Overall, the findings indicate that RoBERTa-based frozen feature extraction constitutes a computationally efficient alternative for hate speech detection on English Twitter data, although further improvements are required to enhance performance in multi-class classification settings.

Keywords: HASOC 2021; Hate Speech; Offensive Content; RoBERTa; Random Forest; Handcrafted Features

1. PENDAHULUAN

Fenomena ujaran kebencian (*hate speech*) dan bahasa ofensif (*offensive language*) di media sosial telah menjadi ancaman serius bagi kohesi sosial dan keamanan ruang digital. *Hate speech* didefinisikan sebagai ungkapan kebencian yang ditujukan kepada individu atau kelompok berdasarkan atribut identitas seperti ras, agama, etnis, gender, atau orientasi seksual, yang berpotensi memicu diskriminasi dan kekerasan [1]. Berbeda dengan hate speech yang umumnya memiliki target, kategori, dan tingkat tertentu terhadap individu atau kelompok, abusive/offensive language merujuk pada bahasa kasar atau menyerang yang tidak selalu berbasis identitas.[2]. Tantangan ini semakin besar pada data media sosial yang banyak mengandung *code-mixing*, kosakata di luar standar baku, singkatan, kesalahan gramatikal, emotikon, dan keterbatasan konteks yang menjadi karakteristik umum teks informal daring [3].

Penelitian terdahulu menunjukkan bahwa pendekatan *machine learning* masih banyak digunakan dalam klasifikasi *hate speech* dan *offensive language* pada media sosial, khususnya Twitter berbahasa Inggris, meskipun tugas ini masih menantang karena ujaran ofensif tidak selalu dapat dikenali hanya dari kemunculan kata kasar, tetapi juga dipengaruhi oleh konteks, target ujaran, serta bentuk ekspresi pengguna. Identifikasi bahasa ofensif di media sosial pada OffensEval-2020 mencakup beberapa subtugas, yaitu identifikasi ujaran ofensif, kategorisasi jenis pelanggaran, dan identifikasi target ujaran [4]. Hal ini menunjukkan bahwa klasifikasi hate speech dan offensive language membutuhkan representasi fitur yang mampu menangkap konteks kalimat secara lebih baik. RoBERTa efektif digunakan dalam deteksi ujaran ofensif

pada Twitter berbahasa Inggris dengan F1-score mencapai 0,9166 pada English Subtask A OffensEval-2020 [5]. Selain itu, kombinasi multilingual RoBERTa dan Random Forest telah digunakan dalam identifikasi hate speech dan offensive content pada HASOC 2020 [6]. TweetEval juga menegaskan pentingnya model bahasa yang sesuai dengan karakteristik data Twitter pada berbagai tugas klasifikasi, termasuk hate speech dan offensive language identification [7].

Berdasarkan temuan tersebut, penelitian ini membutuhkan pipeline yang mampu memanfaatkan representasi kontekstual Transformer tanpa harus bergantung pada proses fine-tuning penuh. Pendekatan frozen feature extractor menjadi relevan karena model pralatih dapat digunakan untuk menghasilkan embedding, sedangkan proses klasifikasi dilakukan oleh algoritma lain yang lebih sederhana Dibandingkan dengan fine-tuning penuh, pendekatan ini memiliki kebutuhan komputasi yang lebih rendah karena proses optimasi hanya dilakukan pada model klasifikasi di tahap akhir, sementara parameter Transformer tetap dibekukan.

Beberapa penelitian melaporkan bahwa strategi ini mampu mempertahankan kualitas representasi semantik yang dihasilkan model pralatih dengan biaya pelatihan yang lebih rendah dibandingkan fine-tuning penuh, terutama pada lingkungan dengan keterbatasan sumber daya komputasi [8]. Pemilihan pendekatan frozen feature extraction dalam penelitian ini juga didasarkan pada pertimbangan efisiensi komputasi, karena proses pelatihan hanya dilakukan pada Random Forest tanpa memperbarui jutaan parameter pada model RoBERTa, sehingga lebih sesuai untuk skenario penelitian dengan sumber daya komputasi yang terbatas[9]. Pembekuan parameter BERT pada tugas klasifikasi teks dapat mempertahankan kualitas representasi sekaligus mengurangi biaya pelatihan/komunikasi [10].

Sejalan dengan itu, kombinasi RoBERTa-large sebagai ekstraksi fitur dan algoritma ensemble sebagai klasifikator efektif untuk deteksi hate speech dan offensive language pada media sosial [11]. Namun, kombinasi RoBERTa sebagai ekstraksi fitur kontekstual dan Random Forest sebagai klasifikator masih perlu dikaji lebih lanjut pada data yang tidak seimbang. Label imbalance merupakan hambatan penting dalam abusive language detection karena dapat menurunkan kemampuan model dalam mengenali kelas minoritas [12]. Oleh karena itu, penelitian ini mengevaluasi pipeline RoBERTa dan Random Forest dengan mempertimbangkan strategi penanganan ketidakseimbangan kelas pada data Twitter berbahasa Inggris.

Pendekatan hibrida menjadi arah yang relevan dalam klasifikasi hate speech dan offensive language karena mampu menggabungkan representasi kontekstual dari Transformer dengan efisiensi algoritma klasifikasi klasik. Studi terbaru menunjukkan bahwa RoBERTa dapat digunakan sebagai ekstraksi fitur dan dikombinasikan dengan algoritma ensemble untuk mendeteksi hate speech serta offensive language pada media sosial[13]. Selain embedding Transformer, hand-crafted features juga relevan sebagai fitur pelengkap karena fitur linguistik eksplisit dapat digabungkan dengan embedding BERT atau XLM-RoBERTa dan diproses menggunakan algoritma machine learning, termasuk Random Forest [14]Namun, dataset hate speech sering menghadapi tantangan berupa cakupan label, kualitas anotasi, dan distribusi data yang perlu dikelola secara hati-hati [15].

Berdasarkan kebutuhan tersebut, Dataset HASOC 2021 Track English dipilih karena menyediakan benchmark berbasis Twitter dengan struktur klasifikasi bertingkat, yaitu Task A sebagai klasifikasi biner HOF dan NOT, serta Task B sebagai klasifikasi lebih rinci ke dalam kelas HATE, OFFN, dan PRFN [16]. Dengan demikian, dataset ini relevan untuk mengevaluasi pipeline hibrida yang menggabungkan embedding RoBERTa, hand-crafted features, Random Oversampling, dan Random Forest sebagai klasifikator.

Random Forest digunakan sebagai klasifikator utama karena sesuai untuk mengolah representasi fitur berdimensi tinggi yang dihasilkan dari embedding RoBERTa dan hand-crafted features. Hal ini didukung oleh penelitian sebelumnya yang menunjukkan bahwa Random Forest dapat bekerja pada data berdimensi tinggi, terutama ketika proses seleksi fitur digunakan untuk memilih subset fitur yang lebih informatif sehingga efisiensi dan performa model dapat meningkat[17]. Dalam pipeline yang diusulkan, embedding RoBERTa berperan sebagai representasi kontekstual untuk menangkap makna kata berdasarkan konteks kalimat, sedangkan hand-crafted features digunakan sebagai fitur pelengkap untuk merepresentasikan informasi eksplisit dari permukaan teks, seperti karakteristik leksikal, pola penggunaan tanda baca, atau struktur sederhana yang muncul pada tweet. Penggunaan fitur buatan ini relevan karena model hibrida yang menggabungkan RoBERTa, Random Forest, dan handcrafted linguistic features terbukti dapat meningkatkan performa klasifikasi pada tugas pemrosesan teks [18].

Selanjutnya, Random Oversampling diterapkan hanya pada data latih agar distribusi kelas minoritas menjadi lebih seimbang sebelum proses klasifikasi. Strategi ini digunakan karena ketidakseimbangan kelas dalam deteksi ujaran abusif dapat menyebabkan model lebih cenderung mengenali kelas mayoritas dan menurunkan performa pada kelas minoritas[12][19]. Dengan demikian, penelitian ini mengevaluasi kontribusi embedding RoBERTa, hand-crafted features, dan Random Oversampling dalam pipeline hibrida berbasis Random Forest untuk klasifikasi hate speech dan offensive language pada Twitter berbahasa Inggris.

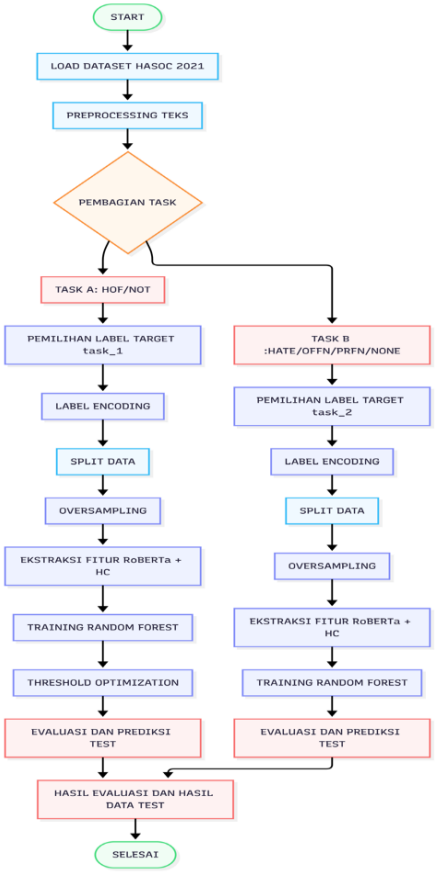
Minimnya kajian yang secara sistematis mengevaluasi pendekatan frozen feature extraction berbasis Transformer dikombinasikan dengan klasifikator ensemble pada dataset hate speech multikelas yang tidak seimbang menjadi landasan utama dilakukannya penelitian ini. Melalui pengujian pipeline RoBERTa frozen feature extractor dan Random Forest pada dataset benchmark HASOC 2021 Track English, penelitian ini tidak hanya mengevaluasi efektivitas konfigurasi dasar, tetapi juga mengukur kontribusi tambahan dari strategi Random Oversampling dan hand-crafted features terhadap performa klasifikasi pada Task A dan Task B secara terpisah. Hasil penelitian ini diharapkan menjadi rujukan empiris bagi pengembangan sistem deteksi hate speech yang efisien secara komputasi, terutama bagi peneliti yang menghadapi keterbatasan infrastruktur dan distribusi kelas yang tidak seimbang sebagai kondisi nyata di lapangan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama sebagaimana ditunjukkan pada bagan alur penelitian. Tahap awal dimulai dengan memuat dataset HASOC 2021 sebagai data utama, kemudian dilakukan preprocessing teks untuk menyesuaikan bentuk tweet agar siap diproses oleh model. Setelah itu, label ditentukan berdasarkan skema task pada dataset, yaitu Task A untuk klasifikasi biner HOF/NOT dan Task B untuk klasifikasi multikelas HATE/OFFN/PRFN/NONE. Data kemudian dibagi menjadi data training dan data validasi agar proses pelatihan dan evaluasi tetap terpisah. Setelah tahap split, alur penelitian dibedakan menjadi dua jalur eksperimen. Pada Task A, data training diseimbangkan melalui oversampling, kemudian dilakukan ekstraksi fitur, pelatihan model Random Forest, optimasi threshold, serta evaluasi dan prediksi akhir. Sementara itu, pada Task B dilakukan penambahan data HASOC 2020 pada data training untuk memperkaya variasi data, dilanjutkan dengan targeted oversampling pada kelas yang jumlahnya lebih rendah, ekstraksi fitur, pelatihan Random Forest, serta evaluasi dan prediksi. Dengan alur ini, penelitian tidak hanya menguji model secara langsung, tetapi juga menerapkan strategi penanganan ketidakseimbangan kelas dan pemisahan evaluasi yang jelas antara Task A dan Task B. Tahapan tersebut dirancang agar model dapat mempelajari pola dari data training secara lebih seimbang, sementara performa akhirnya tetap diukur pada data evaluasi yang tidak ikut dimodifikasi selama proses pelatihan.



Gambar 1. Tahapan Penelitian

2.2 Dataset

Dataset yang digunakan dalam penelitian ini berasal dari shared task HASOC 2021, yaitu kompetisi deteksi ujaran kebencian dan konten ofensif yang diselenggarakan dalam Forum for Information Retrieval Evaluation (FIRE). HASOC 2021 menyediakan dataset berbasis Twitter untuk mengevaluasi sistem klasifikasi ujaran kebencian dan konten ofensif pada beberapa bahasa, termasuk bahasa Inggris[16]Pada penelitian ini, dataset digunakan dalam dua skema klasifikasi, yaitu Task A dan Task B. Task A bertujuan mengklasifikasikan tweet ke dalam dua kategori, yaitu HOF (Hate and Offensive) dan NOT (Not Hate or Offensive), sedangkan Task B digunakan untuk klasifikasi yang lebih rinci terhadap jenis konten bermasalah, seperti HATE, OFFN, dan PRFN. Pada data yang digunakan dalam penelitian ini, label NONE digunakan untuk menunjukkan tweet yang tidak mengandung konten negatif. Selain HASOC 2021 sebagai data utama, penelitian ini juga menggunakan HASOC 2020 sebagai data tambahan karena masih berada pada domain yang sama, yaitu deteksi ujaran kebencian dan konten ofensif pada media sosial [20]. Kedua dataset tersebut berisi tweet berbahasa Inggris yang telah melalui proses anotasi, sehingga dapat digunakan sebagai data berlabel untuk pelatihan dan evaluasi model klasifikasi.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Tabel 1. Dataset Penelitian

Dataset	Jenis	Total	Distribusi Kelas Task A		Distribusi Kelas Task B			
			HOF	NOT	HATE	OFFN	PRFN	NONE
HASOC 2021	Train	3.843	2.501	1.342	683	622	1.196	1.342
HASOC 2020	Train Tambahan	3.708	1.856	1.852	158	321	1.377	1.852
HASOC 2021	Test	1.281	-	-	-	-	-	-

Data HASOC 2021 digunakan sebagai data utama penelitian karena dataset ini dikembangkan untuk deteksi ujaran kebencian dan konten ofensif berbasis Twitter, serta menyediakan skema klasifikasi Task A dan Task B. Data utama tersebut dibagi menjadi 80% data training dan 20% data validasi, karena rasio 80:20 merupakan salah satu rasio yang umum digunakan dalam pembagian data training dan testing pada pemodelan machine learning [21]. Data training digunakan untuk melatih model, sedangkan data validasi digunakan untuk mengevaluasi performa model pada data yang tidak digunakan selama proses pelatihan, sehingga kemampuan generalisasi model dapat diamati dan risiko overfitting dapat dikurangi [22]. Dataset HASOC 2020 digunakan sebagai data tambahan pada data training karena masih berada pada domain yang sama, yaitu deteksi ujaran kebencian dan konten ofensif pada media sosial. Data tambahan ini tidak dimasukkan ke data validasi agar proses evaluasi tetap berfokus pada data utama, yaitu HASOC 2021.

Tabel 2. Contoh Sampel Dataset HASOC 2021

Tweet	Task A	Task B
@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
@ndtv Shameless PM. What else can we say? #ShameOnModi #Resign_PM_Modi #ResignPMmodi	HOF	OFFN
@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. #BengalBurning	NOT	NONE
@Chahal_Shekhar Sorry we won't, why can't your raise your voice for thousands of people who died due to bed crisis and oxygen. Are you people trying to divert this situation and saving #Modi? #Resign_PM_Modi this time it won't work out and even bhakths	HOF	HATE

Berdasarkan Tabel 2, contoh sampel dataset HASOC 2021 ditampilkan untuk menunjukkan bentuk data yang digunakan dalam penelitian. Tweet pada dataset ini memiliki karakteristik khas media sosial, seperti mention, hashtag, URL, bahasa tidak baku, serta ungkapan yang dapat bersifat kasar, menyerang, atau netral. Label yang digunakan mengacu pada label yang tersedia dalam dataset, sehingga peneliti tidak melakukan pelabelan ulang terhadap data. Pada Task A, tweet diklasifikasikan ke dalam dua kategori, yaitu HOF (Hate and Offensive) dan NOT (Not Hate or Offensive), sedangkan pada Task B tweet diklasifikasikan secara lebih rinci ke dalam kategori HATE, OFFN, dan PRFN sesuai skema fine-grained classification pada HASOC 2021. Pada data penelitian ini, label NONE digunakan untuk menunjukkan tweet yang tidak termasuk dalam kategori konten negatif. Dengan demikian, Task A digunakan untuk mengidentifikasi apakah tweet termasuk konten bermasalah atau tidak, sedangkan Task B digunakan untuk menentukan jenis konten tersebut secara lebih spesifik.

2.3 Text Preprocessing

Text preprocessing adalah tahap penting yang bertujuan membersihkan dan menyesuaikan format teks mentah agar dapat diproses secara optimal oleh model yang digunakan Text preprocessing merupakan tahap penting dalam pemrosesan bahasa alami yang bertujuan untuk membersihkan dan menyeragamkan format teks sebelum dilakukan ekstraksi fitur.

2.3.1 Text Preprocessing Roberta

Pada penelitian ini, text preprocessing dilakukan secara minimal dengan mengganti mention pengguna menjadi @user dan tautan URL menjadi http. Pemilihan preprocessing ini didasarkan pada dokumentasi resmi model cardiffnlp/twitter-roberta-base-offensive, karena model tersebut menggunakan format input Twitter dengan penggantian username dan link menjadi placeholder @user dan http [23]. Model ini merupakan Twitter-RoBERTa yang dilatih pada korpus Twitter berukuran besar dan di-fine-tuning untuk offensive language identification menggunakan benchmark TweetEval[24]. Pada penelitian ini, tahapan preprocessing Roberta terdiri dari:

a. *Username Replacement* adalah proses penggantian setiap token yang diawali karakter @ dengan token generik @user. Penggantian dilakukan untuk menghilangkan identitas pengguna yang tidak berkontribusi pada makna semantik tweet sekaligus menjaga privasi. Token @user dipilih sebagai pengganti bukan dihapus sepenuhnya agar struktur dan koherensi kalimat tetap terjaga sebagai input model.
b. *URL Replacement* adalah proses penggantian setiap token yang diawali dengan http menjadi token generik http. Kehadiran tautan eksternal dipertahankan sebagai sinyal kontekstual bahwa tweet menyertakan referensi, namun variasi URL spesifik yang tidak membawa informasi semantik untuk klasifikasi dihilangkan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Tabel 3. Hasil Text Preprocessing untuk Input RoBERTa

Proses	Teks Sebelum Proses	Teks Sesudah Proses
Username Replacement	The report on @TheLeadCNN that @clarissaward filed from India regarding the dire COVID catastrophe was absolutely riveting, devastatingly sad...and some of the best and most important reporting you will see. Do not look away. #IndiaCovidCrisis https://t.co/oHsnIXlEla	The report on @user that @user filed from India regarding the dire COVID catastrophe was absolutely riveting, devastatingly sad...and some of the best and most important reporting you will see. Do not look away. #IndiaCovidCrisis https://t.co/oHsnIXlEla
URL Replacement	The report on @user that @user filed from India regarding the dire COVID catastrophe was absolutely riveting, devastatingly sad...and some of the best and most important reporting you will see. Do not look away. #IndiaCovidCrisis https://t.co/oHsnIXlEla	The report on @user that @user filed from India regarding the dire COVID catastrophe was absolutely riveting, devastatingly sad...and some of the best and most important reporting you will see. Do not look away. #IndiaCovidCrisis http

Pendekatan ini juga sejalan dengan pipeline TweetEval yang memperlakukan mention pengguna dan tautan web secara khusus, yaitu mention dianonimkan dan tautan web dihapus pada proses penyamaan daa, sehingga preprocessing minimal dipilih agar variasi username dan URL yang tidak menjadi fokus klasifikasi dapat dikurangi, tetapi informasi penting seperti hashtag, kata kasar, struktur kalimat, dan konteks tweet tetap dipertahankan[25]

2.3.2 Text Preprocessing TF-IDF

Pada penelitian ini, tahapan preprocessing TF-IDF terdiri dari:

a. Case Folding Case folding dilakukan dengan mengubah seluruh huruf menjadi huruf kecil. Tahap ini bertujuan menyeragamkan bentuk kata agar kata yang sama tidak dianggap sebagai term berbeda hanya karena perbedaan huruf kapital, misalnya Hate, HATE, dan hate [26].
b. Text Cleaning Text cleaning dilakukan dengan menghapus URL, mention pengguna, angka, dan tanda baca. Simbol hashtag (#) dihapus, tetapi kata di dalam hashtag tetap dipertahankan karena dapat memuat informasi topik atau konteks tweet. Contohnya, #ResignModi diubah menjadi resignmodi, bukan dihapus seluruhnya.
c. Stopword Removal Stopword removal dilakukan dengan menghapus kata-kata umum yang tidak terlalu informatif untuk pembobotan TF-IDF. Tahap ini digunakan agar fitur lebih berfokus pada kata yang berpotensi membedakan kelas. Penghapusan stopword juga umum digunakan dalam preprocessing teks untuk mempertahankan kata yang lebih bermakna dalam analisis [27].
d. Penghapusan Token Pendek Token yang terlalu pendek, yaitu token dengan panjang satu karakter, dihapus karena cenderung tidak memberikan informasi klasifikasi yang kuat. Tahap ini membantu mengurangi noise pada vocabulary TF-IDF. Pada penelitian ini, stemming tidak digunakan karena tidak terdapat proses pengubahan kata ke bentuk dasar dalam coding, sehingga bentuk asli kata pada tweet tetap dipertahankan setelah proses cleaning dan stopword removal.

Tabel 4. Hasil Text Preprocessing untuk Input TF-IDF

Proses	Teks Sebelum Proses	Teks Sesudah Proses
Teks asli	@BJP4India @narendramodi @drharshvardhan @AshwiniKChoubey @MoHFW_INDIA Such a shameless tweet #Resign_PM_Modi #ModiResign #ModiKaVaccineJumla #ModiFailsIndia https://t.co/gdfM77oQ40	@BJP4India @narendramodi @drharshvardhan @AshwiniKChoubey @MoHFW_INDIA Such a shameless tweet #Resign_PM_Modi #ModiResign #ModiKaVaccineJumla #ModiFailsIndia https://t.co/gdfM77oQ40
Case folding	@BJP4India @narendramodi @drharshvardhan @AshwiniKChoubey @MoHFW_INDIA Such a shameless tweet #Resign_PM_Modi #ModiResign #ModiKaVaccineJumla #ModiFailsIndia https://t.co/gdfM77oQ40	@bjp4india @narendramodi @drharshvardhan @ashwinikchoubey @mohfw_india such a shameless tweet #resign_pm_modi #modiresign #modikavaccinejumla #modifailsindia https://t.co/gdfm77oq40
Text cleaning	@bjp4india @narendramodi @drharshvardhan @ashwinikchoubey @mohfw_india such a shameless tweet #resign_pm_modi #modiresign #modikavaccinejumla #modifailsindia https://t.co/gdfm77oq40	such a shameless tweet resignpmmodi modiresign modikavaccinejumla modifailsindia
Stopword removal dan token pendek	such a shameless tweet resignpmmodi modiresign modikavaccinejumla modifailsindia	shameless tweet resignpmmodi modiresign modikavaccinejumla modifailsindia

Tabel 4 menunjukkan bahwa preprocessing TF-IDF pada penelitian ini tidak hanya mengganti mention dan URL seperti pada RoBERTa, tetapi membersihkan teks lebih lengkap agar fitur yang dihitung berfokus pada term yang relevan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Setelah proses cleaning, mention dan URL dihapus, simbol hashtag dihilangkan, angka serta tanda baca dibersihkan, sedangkan kata yang berada di dalam hashtag tetap dipertahankan karena masih dapat memuat informasi konteks. Selanjutnya, stopword dan token yang terlalu pendek dihapus untuk mengurangi noise pada vocabulary TF-IDF. Dengan tahapan tersebut, teks yang masuk ke proses pembobotan TF-IDF menjadi lebih ringkas dan lebih berfokus pada kata-kata yang berpotensi membedakan kelas.

2.4 Feature Extraction

2.4.1 RoBERTa Feature Extraction

RoBERTa (Robustly Optimized BERT Pretraining Approach) merupakan model bahasa berbasis Transformer yang dikembangkan dari BERT dengan memperbaiki strategi pre-training, antara lain menghapus objective Next Sentence Prediction, menggunakan dynamic masking, serta melatih model dengan data, batch, dan durasi pelatihan yang lebih besar[28]. Pada penelitian ini, RoBERTa digunakan sebagai feature extractor untuk mengubah teks tweet menjadi representasi numerik berbasis konteks sebelum diklasifikasikan menggunakan Random Forest. Model yang digunakan adalah cardiffnlp/twitter-roberta-base-offensive melalui library Hugging Face Transformers. Model tersebut dipilih karena merupakan Twitter-RoBERTa berbasis RoBERTa-base yang dilatih pada sekitar 58 juta tweet dan di-fine-tuning untuk offensive language identification menggunakan benchmark TweetEval [23].

2.4.2 TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) merupakan metode pembobotan term yang mengubah teks atau tweet menjadi representasi numerik berdasarkan tingkat kepentingan kata dalam dokumen. Metode ini menggabungkan Term Frequency (TF), yaitu frekuensi kemunculan term dalam dokumen, dan Inverse Document Frequency (IDF), yaitu ukuran seberapa jarang term muncul pada keseluruhan dokumen [[29]. Dengan demikian, kata yang sering muncul pada suatu tweet tetapi jarang muncul pada tweet lain akan memiliki bobot lebih tinggi, sedangkan kata yang terlalu umum memperoleh bobot lebih rendah. Pada penelitian ini, TF-IDF digunakan sebagai fitur pembanding terhadap fitur kontekstual RoBERTa karena sederhana, transparan, dan umum digunakan dalam klasifikasi teks. Representasi teks dibentuk menggunakan pendekatan n-gram agar fitur dapat menangkap kata tunggal maupun kombinasi kata atau karakter yang berdekatan [30]. Implementasi TF-IDF dilakukan menggunakan TfidfVectorizer dengan pembobotan frekuensi berbasis logaritmik, IDF smoothing, dan normalisasi vektor default agar representasi dokumen lebih stabil dalam proses klasifikasi.

2.5 Handcrafted Features

Handcrafted features merupakan fitur tambahan yang dirancang secara eksplisit dari karakteristik linguistik dan struktural teks media sosial, seperti panjang teks, huruf kapital berlebihan, tanda baca, mention, URL, hashtag, repetisi karakter, dan kata-kata ofensif. Fitur ini digunakan untuk melengkapi representasi kontekstual RoBERTa karena meskipun model Transformer mampu menangkap makna semantik melalui *self-attention*, beberapa sinyal permukaan tweet yang bersifat eksplisit tetap dapat menjadi indikator penting dalam deteksi *hate speech* dan *offensive content*. Penggunaan *handcrafted features* juga relevan dengan penelitian terdahulu yang menunjukkan bahwa kombinasi fitur tekstual, leksikal, dan karakteristik media sosial dapat membantu klasifikasi ujaran kebencian [31]. Karena dataset HASOC 2021 berasal dari Twitter dan berfokus pada identifikasi *hate speech* serta *offensive content*, penggunaan fitur seperti mention, hashtag, URL, tanda baca, dan kata ofensif dinilai sesuai dengan karakteristik data penelitian.

Tabel 5. Handcrafted Features yang Diekstraksi dari Tweet

No	Fitur	Definisi
1	Text Length	Jumlah seluruh karakter dalam tweet
2	Word Count	Jumlah seluruh kata dalam tweet
3	Average Word Length	Total panjang kata ÷ jumlah kata
4	Uppercase Ratio	Jumlah huruf kapital ÷ jumlah huruf alfabet
5	Caps Word Ratio	Jumlah kata kapital penuh ÷ jumlah kata
6	Exclamation Count	Jumlah tanda seru (!)
7	Question Count	Jumlah tanda tanya (?)
8	Mention Count	Jumlah mention (@user)
9	URL Count	Jumlah URL
10	Special Character Ratio	Jumlah karakter khusus ÷ jumlah karakter
11	Hashtag Count	Jumlah hashtag (#)
12	Offensive Score	Jumlah kata yang ditemukan pada kamus ofensif

Berdasarkan Tabel 5, handcrafted features yang digunakan pada Task A dan Task B mencakup fitur berbasis frekuensi dan fitur berbasis rasio untuk merepresentasikan karakteristik linguistik maupun struktural tweet. Fitur berbasis frekuensi dihitung berdasarkan jumlah kemunculan elemen tertentu dalam teks, sedangkan fitur berbasis rasio diperoleh dari perbandingan antara karakteristik yang diamati dengan jumlah karakter atau kata yang relevan. Selain itu, Offensive Score digunakan untuk merepresentasikan keberadaan kosakata ofensif berdasarkan leksikon yang digunakan dalam

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

penelitian ini. Penyajian definisi dan mekanisme perhitungan setiap fitur bertujuan untuk memberikan deskripsi yang lebih terstruktur mengenai proses ekstraksi fitur yang diterapkan.

2.6 Oversampling

Oversampling merupakan teknik penyeimbangan data yang dilakukan dengan menambah jumlah sampel pada kelas minoritas agar distribusi data training menjadi lebih proporsional. Teknik ini digunakan karena data tidak seimbang dapat membuat classifier cenderung mempelajari kelas mayoritas, sehingga kelas minoritas berisiko kurang terwakili dalam proses pelatihan [32]. Pada penelitian ini, oversampling diterapkan hanya pada data training, sedangkan data validasi dan data test tetap dipertahankan pada distribusi aslinya. Pemisahan ini dilakukan agar evaluasi tetap objektif dan menghindari data leakage, karena penerapan resampling pada seluruh data sebelum evaluasi dapat menyebabkan performa model menjadi bias [33]. Jenis oversampling yang digunakan adalah Random Oversampling (ROS), yaitu penambahan sampel kelas minoritas dengan mengambil ulang sampel yang sudah ada secara acak hingga mencapai jumlah target tertentu. Pemilihan Random Oversampling (ROS) dalam penelitian ini didasarkan pada karakteristik data berupa embedding RoBERTa berdimensi tinggi. Teknik sintetik seperti SMOTE dan ADASYN bekerja dengan membangkitkan sampel baru melalui interpolasi antar titik data, namun pada representasi embedding berdimensi tinggi proses tersebut berpotensi menghasilkan sampel yang kurang merepresentasikan distribusi semantik asli[34]. Namun, pada ruang fitur berdimensi tinggi, proses pembangkitan sampel sintetis dapat menghadapi tantangan dalam merepresentasikan distribusi data asli secara akurat[35]. Oleh karena itu, penelitian ini menggunakan ROS sebagai pendekatan yang lebih sederhana dengan mempertahankan karakteristik data asli. Untuk meminimalkan risiko overfitting akibat duplikasi sampel, proses oversampling hanya diterapkan pada data training, sedangkan data validasi dan data test tetap menggunakan distribusi asli sehingga evaluasi model dilakukan pada data yang tidak mengalami proses resampling.

2.7 Klasifikasi Random Forest

Random Forest merupakan algoritma klasifikasi berbasis ensemble learning yang membangun sejumlah decision tree dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan keputusan akhir yang lebih stabil. Berbeda dari decision tree tunggal yang lebih rentan terhadap perubahan data, Random Forest menggunakan pendekatan bootstrap aggregating dan pemilihan fitur secara acak pada proses pembentukan pohon, sehingga dapat mengurangi overfitting dan meningkatkan kemampuan generalisasi model [36]. Pada penelitian ini, Random Forest digunakan sebagai model klasifikasi setelah tweet diubah menjadi representasi numerik melalui ekstraksi fitur. Implementasi ini menggunakan RandomForestClassifier dari library scikit-learn, yang menyediakan sejumlah parameter untuk mengatur proses pembentukan model, seperti jumlah pohon, kedalaman pohon, jumlah minimum sampel pada *node*, jumlah fitur yang dipertimbangkan pada pemisahan *node*, bobot kelas, serta *criterion* untuk mengukur kualitas pemisahan data . Pada tahap optimasi, beberapa parameter diuji untuk memperoleh konfigurasi yang memberikan performa terbaik berdasarkan data validasi, karena performa Random Forest dapat dipengaruhi secara signifikan oleh pemilihan *hyperparameter* seperti jumlah pohon, kedalaman maksimum, jumlah fitur pada setiap *split*, dan aturan pemisahan *node* [37].

2.8 Parameter Tuning Random Forest

Hyperparameter tuning merupakan proses pencarian kombinasi parameter terbaik agar model dapat menghasilkan performa yang optimal pada data validasi. Performa Random Forest dapat dipengaruhi oleh pengaturan hyperparameter seperti jumlah pohon, jumlah fitur pada setiap split, ukuran node, aturan split, serta strategi sampling [3]. Berikut parameter yang akan diuji :

a. n_estimators : Parameter ini menentukan jumlah decision tree yang dibangun dalam Random Forest. Semakin banyak pohon, prediksi model dapat menjadi lebih stabil, tetapi waktu komputasi juga meningkat. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah 100, 200, 300, 400, 500, 600, 800, 1000, dan 1200.
b. max_depth : Parameter ini menentukan kedalaman maksimum setiap pohon. Kedalaman yang terlalu kecil dapat membuat model kurang menangkap pola data, sedangkan kedalaman yang terlalu besar dapat meningkatkan risiko overfitting. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah 20, 30, 40, 50, 100, dan None. Nilai None berarti kedalaman pohon tidak dibatasi secara eksplisit.
c. min_samples_split : Parameter ini menentukan jumlah minimum sampel yang diperlukan agar sebuah node dapat dibagi lagi. Nilai yang lebih besar membuat pohon lebih berhati-hati dalam membentuk cabang baru. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah 2, 3, 5, 8, dan 10.
d. min_samples_leaf : Parameter ini menentukan jumlah minimum sampel yang harus berada pada leaf node atau node akhir. Parameter ini membantu mencegah model membuat keputusan dari jumlah sampel yang terlalu kecil. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah 1, 2, 3, 4, 5, dan 10.
e. max_features : Parameter ini menentukan jumlah fitur yang dipertimbangkan saat proses pemisahan node. Pada Random Forest, tidak semua fitur selalu digunakan pada setiap split agar pohon-pohon yang terbentuk memiliki variasi keputusan. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah sqrt, log2, 0.2, 0.3, 0.4, dan 0.5.
f. max_samples : Parameter ini menentukan proporsi sampel training yang digunakan untuk membentuk setiap pohon ketika bootstrap aktif. Parameter ini digunakan pada beberapa skenario RoBERTa untuk mengatur variasi data antar pohon. Pada seluruh skenario penelitian ini, nilai yang diuji atau digunakan adalah 0.6, 0.7, 0.8, 0.9, dan None.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

g. criterion : Parameter ini menentukan ukuran yang digunakan untuk menilai kualitas pemisahan node pada decision tree. Pada penelitian ini, nilai yang diuji atau digunakan adalah gini dan entropy. Keduanya tidak digunakan bersamaan dalam satu model, tetapi menjadi alternatif criterion pada skenario eksperimen. gini mengukur ketidakmurnian kelas pada node, sedangkan entropy mengukur ketidakpastian distribusi kelas [3].

Setelah seluruh kombinasi parameter diuji, konfigurasi terbaik dipilih berdasarkan nilai F1-macro. Proses pencarian hyperparameter dilakukan menggunakan RandomizedSearchCV dengan Stratified K-Fold Cross Validation sebanyak 3 fold. Pendekatan stratified digunakan untuk mempertahankan proporsi distribusi kelas pada setiap fold mengingat dataset memiliki ketidakseimbangan kelas[38]. Dengan demikian, pemilihan hyperparameter tidak hanya bergantung pada satu pembagian data, tetapi mempertimbangkan rata-rata performa dari beberapa fold selama proses optimasi. F1-macro digunakan sebagai metrik optimasi karena penelitian ini melibatkan distribusi kelas yang tidak seimbang, sehingga evaluasi tidak hanya bergantung pada akurasi, tetapi juga memperhatikan performa setiap kelas. Hyperparameter terbaik yang diperoleh dari proses cross-validation kemudian digunakan untuk membangun model akhir dan dievaluasi pada data validasi yang dipisahkan sebelumnya dengan rasio 80:20.

2.9 Optimasi

Optimasi model dilakukan untuk meningkatkan performa klasifikasi pada dataset HASOC 2021 yang memiliki karakteristik teks Twitter, distribusi kelas tidak seimbang, serta kategori label yang berbeda antara Task A dan Task B. Model terbaik dipilih berdasarkan F1-macro, bukan hanya akurasi, karena dataset yang tidak seimbang dapat membuat akurasi tampak tinggi tetapi belum tentu merepresentasikan performa setiap kelas secara adil [39] Tahapan optimasi pada penelitian ini adalah sebagai berikut :

a. Penambahan data training: Pada beberapa skenario Task B, dataset HASOC 2020 ditambahkan hanya ke data training untuk memperkaya variasi pola teks, terutama pada kelas yang jumlah datanya lebih sedikit. Data validasi tetap berasal dari HASOC 2021 agar evaluasi tetap berfokus pada data utama penelitian.
b. Pengujian representasi fitur : RoBERTa digunakan sebagai fitur utama karena menghasilkan representasi kontekstual dari tweet. Beberapa skenario menambahkan handcrafted features untuk menangkap karakteristik eksplisit, seperti huruf kapital, tanda seru, hashtag, mention, URL, dan kata ofensif. TF-IDF digunakan sebagai pembanding karena metode ini merepresentasikan teks berdasarkan bobot kata dan n-gram yang umum digunakan dalam klasifikasi teks[40] .
c. Penyeimbangan kelas pada data training : Oversampling diterapkan hanya pada data training. Pada Task A, oversampling digunakan untuk membantu keseimbangan kelas biner HOF dan NOT. Pada Task B, oversampling dilakukan secara terarah pada kelas HATE dan OFFN karena jumlahnya lebih rendah dibandingkan kelas lain. Data validasi dan test tidak di-oversampling agar hasil evaluasi tidak bias.
d. Hyperparameter tuning Random Forest : Random Forest dioptimasi dengan menguji beberapa parameter, seperti n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, max_samples, class_weight, dan criterion. Parameter gini dan entropy digunakan sebagai alternatif criterion pada proses tuning, bukan digunakan bersamaan dalam satu model.
e. Threshold tuning : Threshold tuning diterapkan pada Task A karena Task A merupakan klasifikasi biner. Nilai threshold tidak selalu ditetapkan 0,50, tetapi diuji pada beberapa nilai probabilitas untuk memperoleh F1-macro terbaik. Pada Task B, threshold tuning tidak digunakan karena klasifikasinya multikelas, sehingga optimasi difokuskan pada kombinasi fitur, oversampling, class weight, dan hyperparameter Random Forest.

Dengan tahapan tersebut, model terbaik tidak dipilih secara acak, tetapi berdasarkan pengujian kombinasi data, fitur, parameter, dan threshold. Fokus utama optimasi tetap pada pipeline RoBERTa-Random Forest, sedangkan TF-IDF digunakan sebagai pembanding untuk menunjukkan posisi fitur kontekstual terhadap fitur statistik berbasis kata
Model terbaik kemudian dipilih berdasarkan nilai F1-macro tertinggi karena metrik ini lebih sesuai untuk dataset dengan distribusi kelas yang tidak seimbang.

2.10 Evaluasi

Evaluasi model dilakukan untuk mengukur kemampuan model dalam mengklasifikasikan tweet pada dataset HASOC 2021. Metrik yang digunakan meliputi accuracy, precision, recall, dan F1-score karena keempat metrik tersebut umum digunakan untuk menilai performa model klasifikasi berdasarkan hasil confusion matrix [41]. Accuracy menunjukkan proporsi prediksi yang benar terhadap seluruh data, tetapi pada dataset dengan distribusi kelas tidak seimbang, accuracy tidak cukup dijadikan satu-satunya dasar penilaian karena dapat lebih dipengaruhi oleh kelas mayoritas . Oleh karena itu, penelitian ini menggunakan F1-macro sebagai metrik utama. F1-score digunakan karena menggabungkan precision dan recall dalam satu ukuran, sehingga mampu menilai keseimbangan antara ketepatan prediksi dan kemampuan model menemukan data pada suatu kelas. Pada klasifikasi multikelas, F1-macro dihitung dengan mengambil rata-rata F1-score dari setiap kelas tanpa mempertimbangkan proporsi jumlah data pada masing-masing kelas, sehingga setiap kelas memiliki kontribusi yang sama dalam evaluasi [3]. Dengan demikian, F1-macro lebih sesuai digunakan pada penelitian ini karena dataset HASOC memiliki distribusi kelas yang tidak seimbang dan penelitian tidak hanya menilai performa pada kelas mayoritas, tetapi juga pada kelas minoritas.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

$$F1_{macro} = \frac{1}{K}\sum_{i=1}^{K} F1_i \quad (2)$$

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, evaluasi model tidak dilakukan melalui submit langsung ke leaderboard resmi HASOC, melainkan melalui perbandingan manual terhadap hasil yang telah dipublikasikan oleh penyelenggara. Pendekatan ini digunakan untuk menempatkan performa model yang diusulkan dalam konteks benchmark HASOC dengan menggunakan metrik evaluasi yang sama, yaitu F1-score, Accuracy, Precision, dan Recall.

3.1 Hasil Eksperimen Metode Dasar (Baseline)

Metode baseline pada penelitian ini digunakan sebagai gambaran awal performa model sebelum dilakukan optimasi. Pada tahap baseline, data yang digunakan adalah HASOC 2021 dengan model RoBERTa sebagai ekstraksi fitur dan Random Forest sebagai klasifier. Model baseline dibuat tanpa penambahan data HASOC 2020, tanpa oversampling, tanpa handcrafted features, tanpa hyperparameter tuning, dan tanpa threshold tuning, sehingga hasilnya dapat menjadi acuan untuk melihat pengaruh optimasi pada tahap berikutnya. Berdasarkan pengujian pada data validasi, baseline Task A memperoleh F1-macro sebesar 0,7884**,** accuracy 0,8127**,** precision macro 0,7987**,** dan recall macro 0,7813**.** Sementara itu, baseline Task B memperoleh F1-macro sebesar 0,5906**,** accuracy 0,6671, precision macro 0,6241**,** dan recall macro 0,5931. Hasil ini menunjukkan bahwa model awal sudah cukup baik pada Task A, tetapi masih lebih rendah pada Task B karena klasifikasi multikelas lebih kompleks dibandingkan klasifikasi biner. Oleh karena itu, hasil baseline digunakan sebagai pembanding untuk menilai peningkatan performa setelah dilakukan optimasi, seperti penambahan data training, oversampling, penambahan fitur, tuning parameter, dan threshold tuning.

3.2 Penelusuran Model Optimal Data Validasi

Penelusuran model optimal dilakukan pada data validasi untuk memilih skenario terbaik sebelum model diterapkan pada data test. Pada Task A, pengujian diarahkan untuk klasifikasi biner HOF dan NOT, sehingga skenario yang dibandingkan meliputi penggunaan fitur RoBERTa dan TF-IDF, penambahan handcrafted features, variasi parameter Random Forest, serta penyesuaian threshold. Pada Task B, pengujian diarahkan untuk klasifikasi multikelas HATE, OFFN, PRFN, dan NONE, sehingga skenario yang dibandingkan mencakup penggunaan fitur RoBERTa dan TF-IDF, handcrafted features, penambahan data training HASOC 2020, serta variasi parameter Random Forest. Threshold tuning hanya diterapkan pada Task A karena klasifikasinya bersifat biner, sedangkan pada Task B prediksi ditentukan berdasarkan kelas dengan probabilitas tertinggi. Hasil pengujian setiap skenario ditampilkan pada Tabel 4 untuk Task A dan Tabel 5 untuk Task B, dengan F1-macro sebagai acuan utama pemilihan model karena metrik tersebut lebih menunjukkan keseimbangan performa antar kelas, sedangkan accuracy digunakan sebagai informasi pendukung performa secara keseluruhan.

Tabel 6. Penelusuran Model Optimal dan Pengujian Terhadap Data Validasi Task A

Metode	Dataset Train	HF Fitur	Best Parameter RF						Score		T	Ket.
			C	m_d	m_sl	n_est	m_ss	m_f	m_s	F1-macro	Accuracy		
RoBERTa	2021	12	entropy	30	3	800	8	0.2	None	0,8048	0,8244	0,61	TA1
RoBERTa	2021	0	entropy	40	1	300	3	0.3	0.8	0,8035	0,8257	0,60	TA2
RoBERTa	2021	12	gini	50	1	800	3	0.5	None	0,7902	0,8101	0,62	TA4
RoBERTa	2021	0	gini	None	1	500	3	0.3	None	0,7896	0,8140	0,59	TA5
TF-IDF	2021	12	gini	None	1	200	2	log2	None	0,7561	0,7750	0,69	TA6
TF-IDF	2021	0	gini	None	1	200	2	log2	None	0,7509	0,7841	0,61	TA7
TF-IDF	2021	0	entropy	None	1	300	5	log2	None	0,7455	0,7724	0,62	TA8

Berdasarkan Tabel 6, performa terbaik pada Task A diperoleh oleh skenario TA1, yaitu kombinasi RoBERTa dengan 12 handcrafted features dan Random Forest menggunakan criterion entropy. Skenario ini menghasilkan F1-macro tertinggi sebesar 0,8048 dengan accuracy 0,8244 pada threshold 0,61. Nilai tersebut menunjukkan bahwa TA1 memberikan hasil paling seimbang dalam mengenali kelas HOF dan NOT dibandingkan skenario lainnya. Skenario TA2 memang memiliki accuracy sedikit lebih tinggi, yaitu 0,8257, tetapi F1-macro-nya berada di bawah TA1, sehingga TA1 tetap dipilih sebagai model optimal karena fokus evaluasi penelitian ini adalah keseimbangan performa antar kelas. Jika dilihat dari jenis fitur, seluruh skenario RoBERTa menghasilkan F1-macro lebih tinggi dibandingkan skenario TF-IDF. TF-IDF terbaik hanya memperoleh F1-macro 0,7561 pada TA6, sehingga dapat disimpulkan bahwa representasi kontekstual RoBERTa lebih efektif untuk Task A. Selain itu, hasil TA1 juga menunjukkan bahwa penambahan handcrafted features masih memberikan kontribusi terhadap performa model, karena fitur tambahan tersebut membantu menangkap karakteristik eksplisit tweet yang tidak selalu cukup direpresentasikan oleh embedding saja. Dengan demikian, model TA1 digunakan sebagai konfigurasi terbaik dari hasil validasi sebelum dilakukan pengujian pada data test.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Tabel 7. Penelusuran Model Optimal dan Pengujian Terhadap Data Validasi Task B

Metode	Dataset Train	HC	Best Parameter RF						Score		Ket.
			C	m_d	m_l	n_est	m_ss	m_f	m_s	F1-macro	Accuracy	
RoBERTa	2021	Tidak	entropy	50	3	400	3	0.3	None	0,6392	0,6788	TB1
RoBERTa	2021	Ya	entropy	50	3	400	3	0.3	None	0,6262	0,6697	TB2
RoBERTa	2021	Tidak	gini	50	3	400	3	0.3	None	0,6121	0,6489	TB3
RoBERTa	2021+ 2020	Ya	entropy	50	3	400	3	0.3	None	0,6076	0,6580	TB4
RoBERTa	2021+ 2020	Ya	entropy	40	1	500	2	sqrt	0.8	0,6020	0,6645	TB5
TF-IDF	2021+ 2020	17	entropy	None	1	400	5	sqrt	None	0,5503	0,6047	TB7
TF-IDF	2021+ 2020	17	gini	None	1	400	5	sqrt	None	0,5388	0,5917	TB8
TF-IDF	2021+ 2020	Tidak	gini	None	1	400	5	sqrt	None	0,5433	0,5852	TB9

Berdasarkan Tabel 7, skenario terbaik pada data validasi Task B diperoleh oleh TB1, yaitu kombinasi RoBERTa tanpa handcrafted features dengan Random Forest ber-criterion entropy. Skenario ini menghasilkan F1-macro tertinggi sebesar 0,6392 dan accuracy 0,6788, sehingga dipilih sebagai model optimal untuk Task B. Hasil ini menunjukkan bahwa pada klasifikasi multikelas, representasi kontekstual RoBERTa tanpa tambahan fitur manual sudah mampu memberikan performa paling seimbang dibandingkan skenario lainnya. Skenario RoBERTa dengan handcrafted features pada TB2 memperoleh F1-macro 0,6262, sedangkan skenario RoBERTa dengan criterion gini pada TB3 memperoleh F1-macro 0,6121, sehingga keduanya masih berada di bawah TB1. Penambahan data HASOC 2020 pada skenario TB4 dan TB5 juga belum meningkatkan performa validasi, yang menunjukkan bahwa penambahan data tidak selalu langsung memperbaiki hasil jika distribusi atau karakteristik datanya berbeda dari data utama. Sementara itu, skenario TF-IDF terbaik terdapat pada TB7 dengan F1-macro 0,5503, masih lebih rendah dibandingkan skenario RoBERTa. Dengan demikian, model TB1 dipilih sebagai konfigurasi terbaik Task B karena memberikan F1-macro tertinggi pada data validasi dan menunjukkan performa paling seimbang untuk klasifikasi HATE, OFFN, PRFN, dan NONE.

3.3 Pengujian Terhadap Data Test

Model optimal yang diperoleh dari proses validasi diterapkan pada data test HASOC 2021 yang belum digunakan pada tahap training maupun validasi. Pengujian dilakukan menggunakan Random Forest sebagai classifier untuk melihat kemampuan model dalam mengklasifikasikan data baru. Data test yang digunakan berjumlah 1.281 tweet, dan hasil pengujiannya ditampilkan pada tabel berikut berdasarkan metrik F1-score, accuracy, precision, dan recall ditampilkan dalam Tabel 6 dan Tabel 7.

Tabel 8. Hasil Pengujian terhadap data test Task A

ID	Nama	Metode	Dataset	F1-Score	Accuracy	Precision	Recall
TA1	Run 1	Random Forest + Roberta	H21	80.25%	79.36%	79.00%	79.98%
TA5	Run 2	Random Forest + Roberta	H21	80.00%	81.19%	79.96%	80.04%
TA8	Run 3	Random Forest + TF-IDF	H21	73.83%	76.03%	74.69%	73.33%

Berdasarkan Tabel 6, hasil pengujian data test Task A menunjukkan bahwa model berbasis RoBERTa masih memberikan performa lebih baik dibandingkan TF-IDF. Skenario model pertama (Run 1) memperoleh F1-Score tertinggi sebesar 80,25%, dengan accuracy 79,36%, precision 79,00%, dan recall 79,98%. Sementara itu, (Run 2) menghasilkan accuracy lebih tinggi, yaitu 81,19%, tetapi F1-Score-nya sedikit lebih rendah, yaitu 80,00%. Karena penelitian ini lebih menekankan keseimbangan performa melalui F1-Score, maka (Run 1) dapat dianggap sebagai hasil terbaik pada pengujian Task A. Adapun skenario (Run 3) yang menggunakan TF-IDF memperoleh F1-Score 73,83%, lebih rendah dibandingkan dua skenario RoBERTa. Hal ini menunjukkan bahwa representasi kontekstual RoBERTa lebih efektif untuk membedakan tweet HOF dan NOT pada data test dibandingkan representasi berbasis bobot kata TF-IDF.

Tabel 9. Hasil Pengujian terhadap data test Task B

ID	Nama	Metode	Data Train	F1-Score	Accuracy	Precision	Recall
TB2	Run 4	Random Forest + Roberta	H21	64.64%	59.44%	60.10%	60.33%
TB1	Run 5	Random Forest + Roberta	H21	60.13%	65.18%	60.42%	60.78%
TB7	Run 6	Random Forest + TF-IDF	H20+H21	56.11%	61.67%	55.96%	56.74%

Berdasarkan Tabel 7, hasil pengujian data test Task B menunjukkan bahwa skenario terbaik diperoleh oleh (Run 4), yaitu Random Forest dengan fitur RoBERTa menggunakan data HASOC 2021. Skenario ini menghasilkan F1-Score

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

tertinggi sebesar 64,64%, dengan accuracy 59,44%, precision 60,10%, dan recall 60,33%. Nilai accuracy pada (Run 5) memang lebih tinggi, yaitu 65,18%, tetapi F1-Score (Run 5) lebih rendah dibandingkan TB2, yaitu 60,13%. Oleh karena itu, (Run 4) lebih tepat dipilih sebagai model terbaik untuk Task B karena mampu memberikan performa yang lebih seimbang pada klasifikasi multikelas. Sementara itu, skenario (Run 6) berbasis TF-IDF memperoleh F1-Score 56,11%, sehingga masih berada di bawah skenario RoBERTa. Hasil ini menunjukkan bahwa pada Task B, penambahan data training HASOC 2020 dapat membantu model RoBERTa-Random Forest mengenali variasi kelas secara lebih baik, terutama karena Task B memiliki kategori yang lebih rinci dibandingkan Task A.

Analisis hasil menunjukkan bahwa performa model pada Task B masih lebih rendah dibandingkan Task A, yang mengindikasikan bahwa klasifikasi multikelas pada label HATE, OFFN, PRFN, dan NONE memiliki tingkat kesulitan yang lebih tinggi. Berdasarkan karakteristik label pada dataset HASOC 2021, potensi kesalahan klasifikasi dapat terjadi pada tweet yang mengandung ekspresi negatif atau kata-kata kasar, tetapi tidak secara tegas merepresentasikan perbedaan antara ujaran kebencian (HATE), bahasa ofensif (OFFN), dan ujaran profan (PRFN). Kondisi tersebut menyebabkan batas antar kelas menjadi relatif berdekatan sehingga meningkatkan ambiguitas klasifikasi. Temuan ini menunjukkan bahwa meskipun kombinasi frozen RoBERTa dan Random Forest mampu menangkap pola umum pada data Twitter, representasi fitur yang digunakan belum sepenuhnya mampu membedakan karakteristik fine-grained antar kategori ofensif secara konsisten. Hal ini tercermin dari penurunan nilai F1-macro pada Task B dibandingkan Task A, yang menunjukkan bahwa pemisahan kategori ofensif yang lebih spesifik masih menjadi tantangan dalam proses klasifikasi.

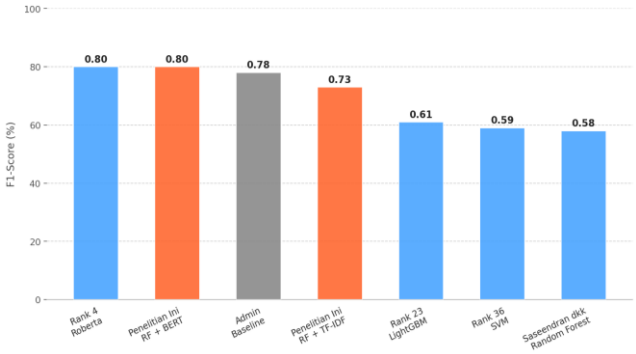
3.4 Ranking LeaderBoard

Pada penelitian ini, leaderboard digunakan sebagai referensi untuk memberikan konteks terhadap performa model yang diusulkan pada benchmark HASOC 2021. Perbandingan dilakukan secara manual dengan mencocokkan nilai F1-score hasil penelitian terhadap hasil yang telah dipublikasikan pada leaderboard dan paper HASOC 2021 [20]. Namun, perbandingan tersebut tidak dapat dianggap sebagai evaluasi langsung karena setiap sistem dapat menggunakan konfigurasi model, strategi pelatihan, sumber data tambahan, maupun sumber daya komputasi yang berbeda. Oleh karena itu, hasil leaderboard dalam penelitian ini digunakan untuk menggambarkan posisi relatif performa model pada benchmark yang sama, bukan untuk menyatakan keunggulan metode secara langsung. Dengan demikian, Tabel 8 dan Tabel 9 menyajikan perbandingan performa model penelitian terhadap hasil yang telah dilaporkan pada HASOC 2021.

Tabel 10. Perbandingan hasil pengujian Task A

Peneliti	Model	F1-score
Rank 4	Roberta	0.80
Rank 23	LightGBM	0.61
Rank 36	SVM	0.59
Saseendran dkk[42]	Random Forest	0.58
Penelitian ini	Random Forest + BERT	0.80
Penelitian ini	Random Forest+TF-IDF	0.73
Admin	Baseline	0.78

Berdasarkan Tabel 8, model Random Forest + RoBERTa memperoleh F1-score sekitar 0,80 pada Task A. Hasil tersebut menunjukkan bahwa representasi kontekstual yang dihasilkan RoBERTa mampu mendukung kinerja Random Forest secara efektif pada tugas klasifikasi hate speech dan offensive language dalam dataset HASOC 2021. Selain itu, performa yang diperoleh lebih tinggi dibandingkan model baseline yang digunakan dalam penelitian ini, yang mengindikasikan bahwa proses optimasi melalui pemilihan fitur, tuning hyperparameter, dan threshold tuning memberikan kontribusi terhadap peningkatan performa model. Perbandingan dengan hasil pada leaderboard HASOC 2021 digunakan sebagai referensi kontekstual untuk menggambarkan posisi relatif performa model pada benchmark yang sama dan perlu diinterpretasikan secara hati-hati karena setiap sistem dapat menggunakan konfigurasi model, strategi pelatihan, maupun sumber daya komputasi yang berbeda.



Gambar 2. Diagram Perbandingan F1-Score Pengujian Task A

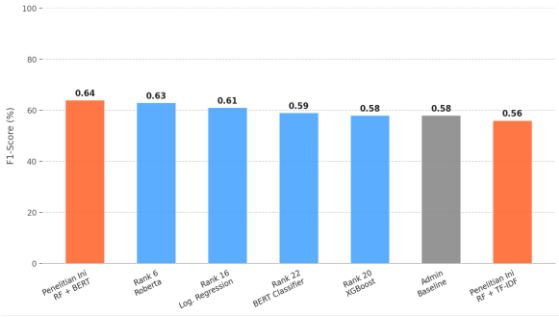
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1388-1402
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1157

Skenario Random Forest + TF-IDF tetap ditampilkan sebagai pembanding konvensional untuk melihat perbedaan performa antara fitur berbasis bobot kata dan fitur kontekstual. Hasil TF-IDF yang lebih rendah dibandingkan RoBERTa menunjukkan bahwa representasi kontekstual lebih mampu menangkap makna tweet pada Task A, terutama dalam membedakan kelas HOF dan NOT. Dengan demikian, Tabel 8 dan Gambar 2 tidak hanya menunjukkan posisi model terhadap hasil penelitian lain, tetapi juga memperlihatkan bahwa optimasi berbasis RoBERTa memberikan performa yang lebih kuat dibandingkan baseline maupun TF-IDF sebagai metode pembanding.

Tabel 11. Perbandingan hasil pengujian Task B

Peneliti	Model	F1-score
Rank 6	Roberta	0.63
Rank 16	Logistic Regression	0.61
Rank 22	BERT classifier	0.59
Rank 20	XGBoost	0.58
Penelitian ini	Random Forest + BERT	0.64
Penelitian ini	Random Forest +TF-IDF	0.56
Admin	Baseline	0.58

Berdasarkan Tabel 9, metode Random Forest + RoBERTa pada penelitian ini memperoleh F1-score sebesar 0,64 pada Task B. Hasil tersebut menunjukkan bahwa pendekatan yang diusulkan masih mampu mempertahankan performa yang memadai pada skenario klasifikasi multikelas yang melibatkan kategori HATE, OFFN, PRFN, dan NONE. Dibandingkan Task A, penurunan performa pada Task B mengindikasikan bahwa pemisahan antar kategori ofensif yang lebih spesifik masih menjadi tantangan bagi model. Kondisi ini menunjukkan bahwa representasi fitur yang berhasil membedakan konten ofensif dan non-ofensif belum tentu mampu memberikan pemisahan yang sama baiknya ketika kategori ofensif dipecah menjadi beberapa kelas yang memiliki karakteristik linguistik yang saling berdekatan. Oleh karena itu, peningkatan kemampuan klasifikasi pada Task B masih memerlukan representasi fitur yang lebih diskriminatif untuk membedakan nuansa antar kategori secara lebih konsisten.



Gambar 3. Diagram Perbandingan F1-Score Pengujian Task B

Sementara itu, metode Random Forest + TF-IDF memperoleh F1-score 0,56, lebih rendah dibandingkan baseline sebesar 0,58. Hal ini menunjukkan bahwa fitur berbasis bobot kata belum cukup kuat untuk menangkap konteks tweet pada klasifikasi multikelas, terutama ketika label yang dibedakan memiliki karakteristik yang berdekatan seperti HATE, OFFN, dan PRFN. Dengan demikian, hasil pada Gambar 3 memperlihatkan bahwa pendekatan berbasis RoBERTa lebih sesuai digunakan pada Task B dibandingkan TF-IDF, karena mampu menghasilkan representasi teks yang lebih kontekstual dan memberikan performa yang lebih stabil pada data uji.

3.5 Pembahasan

Dari hasil penelitian, temuan tersebut sejalan dengan hasil OffensEval-2020 [4] yang menunjukkan bahwa model berbasis RoBERTa mampu memberikan performa tinggi pada tugas offensive language identification karena kemampuannya dalam menghasilkan representasi semantik yang lebih kaya dibandingkan pendekatan berbasis fitur tradisional. Hasil penelitian ini juga mendukung temuan HASOC 2020 [6] yang menunjukkan bahwa representasi fitur berbasis Transformer dapat dimanfaatkan secara efektif sebagai masukan bagi algoritma machine learning konvensional untuk deteksi hate speech dan offensive content. Pada penelitian ini, keunggulan tersebut tercermin dari performa RoBERTa yang secara konsisten lebih tinggi dibandingkan TF-IDF pada Task A maupun Task B. Dengan demikian, hasil penelitian ini memperlihatkan bahwa pendekatan frozen feature extraction masih dapat digunakan sebagai alternatif yang relevan untuk memanfaatkan kemampuan representasi Transformer dengan kebutuhan komputasi yang lebih rendah. Selain itu, hasil eksperimen menunjukkan bahwa kontribusi fitur tambahan tidak selalu memberikan peningkatan performa yang konsisten pada seluruh skenario. Pada Task A, penambahan handcrafted features menghasilkan peningkatan performa dibandingkan penggunaan embedding RoBERTa saja. Namun, pada Task B peningkatan yang sama tidak selalu terjadi. Temuan ini mengindikasikan bahwa fitur-fitur eksplisit seperti panjang teks, penggunaan huruf kapital, tanda baca, hashtag, mention, URL, dan kata ofensif lebih efektif untuk membantu

membedakan konten ofensif dan non-ofensif dibandingkan untuk membedakan kategori ofensif yang lebih spesifik. Perbedaan performa antara Task A dan Task B juga memberikan gambaran mengenai karakteristik masing-masing tugas klasifikasi. Pada Task A, model hanya perlu membedakan antara konten bermasalah dan tidak bermasalah sehingga batas antar kelas relatif lebih jelas. Sebaliknya, pada Task B model harus membedakan kategori HATE, OFFN, PRFN, dan NONE yang memiliki karakteristik linguistik yang saling berdekatan. Kondisi tersebut menyebabkan representasi fitur yang efektif untuk mendeteksi keberadaan konten ofensif belum tentu mampu memisahkan jenis konten ofensif secara konsisten. Temuan ini juga konsisten dengan laporan HASOC 2021 [16] yang menunjukkan bahwa klasifikasi hate speech secara fine-grained masih menjadi tantangan karena adanya kemiripan karakteristik linguistik antar kategori serta ambiguitas konteks pada teks media sosial. Pada penelitian ini, kondisi tersebut tercermin dari performa Task B yang lebih rendah dibandingkan Task A, yang mengindikasikan bahwa pemisahan kategori HATE, OFFN, dan PRFN masih memerlukan representasi fitur yang lebih diskriminatif. Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi representasi kontekstual RoBERTa dan Random Forest mampu memberikan performa yang baik pada klasifikasi hate speech dan offensive language. Temuan ini memperkuat pandangan bahwa pendekatan hibrida yang menggabungkan embedding Transformer dengan algoritma machine learning konvensional masih memiliki potensi untuk digunakan sebagai alternatif yang efisien secara komputasi. Namun demikian, kemampuan model dalam membedakan kategori ofensif yang lebih rinci masih dapat ditingkatkan melalui pengembangan representasi fitur yang lebih diskriminatif maupun pendekatan fine-tuning yang lebih spesifik terhadap karakteristik dataset.

4. KESIMPULAN

Penelitian ini membuktikan bahwa kombinasi RoBERTa sebagai ekstraksi fitur kontekstual dan Random Forest sebagai model klasifikasi mampu meningkatkan performa deteksi ujaran kebencian dan konten ofensif pada dataset HASOC 2021. Kontribusi utama penelitian ini terletak pada penelusuran beberapa skenario optimasi, bukan hanya menjalankan satu model secara langsung. Optimasi dilakukan melalui penggunaan preprocessing yang sesuai dengan karakteristik model, pengujian fitur RoBERTa dengan dan tanpa handcrafted features, penerapan oversampling pada data training, penambahan data HASOC 2020 pada skenario Task B, tuning parameter Random Forest, serta threshold tuning pada Task A. Hasil eksperimen menunjukkan bahwa pendekatan berbasis RoBERTa lebih unggul dibandingkan TF-IDF sebagai pembanding konvensional. Pada Task A, model terbaik diperoleh dari kombinasi Random Forest + RoBERTa + handcrafted features, dengan F1-score data test sebesar 80,25%, lebih tinggi dibandingkan TF-IDF terbaik yang memperoleh 73,83%. Pada Task B, model terbaik diperoleh dari Random Forest + RoBERTa dengan data HASOC 2021, yang menghasilkan F1-score sebesar 64,64%, sedangkan TF-IDF terbaik memperoleh 56,11%. Hasil ini menunjukkan bahwa RoBERTa lebih mampu menangkap konteks tweet dibandingkan TF-IDF yang hanya bergantung pada bobot kemunculan kata. Selain menunjukkan keunggulan representasi kontekstual RoBERTa dibandingkan TF-IDF, hasil penelitian ini mengindikasikan bahwa pendekatan frozen feature extraction tetap mampu menghasilkan representasi fitur yang informatif ketika dikombinasikan dengan algoritma machine learning konvensional seperti Random Forest. Temuan ini menunjukkan bahwa pemanfaatan embedding Transformer tidak selalu harus dilakukan melalui fine-tuning penuh untuk memperoleh performa yang memadai pada tugas klasifikasi hate speech dan offensive language.Hasil eksperimen juga menunjukkan bahwa kontribusi fitur tambahan maupun strategi penanganan ketidakseimbangan kelas tidak selalu memberikan peningkatan performa yang konsisten pada seluruh skenario. Efektivitas suatu konfigurasi model tidak hanya ditentukan oleh kompleksitas metode yang digunakan, tetapi juga oleh karakteristik data dan tugas klasifikasi yang dihadapi. Dengan demikian, penelitian ini menunjukkan bahwa pendekatan hibrida yang mengombinasikan representasi kontekstual dari Transformer dan algoritma machine learning konvensional masih memiliki potensi untuk digunakan sebagai alternatif yang efisien pada klasifikasi hate speech dan offensive language, khususnya ketika ketersediaan sumber daya komputasi menjadi salah satu pertimbangan dalam pengembangan model. Penelitian ini berfokus pada evaluasi pendekatan frozen feature extraction sehingga tidak mencakup perbandingan langsung dengan skenario fine-tuning end-to-end. Selain itu, eksplorasi strategi penanganan ketidakseimbangan kelas yang lebih beragam serta penggunaan klasifikator lain juga berpotensi meningkatkan performa, khususnya pada klasifikasi fine-grained di Task B.

REFERENCES

[1] S. Dharmawan, V. C. Mawardi, and N. J. Perdana, "Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT)," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 11, no. 1, Jun. 2023, doi: 10.24912/jiksi.v11i1.24066.

[2] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2023.e18647.

[3] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.01406125.

[4] M. Zampieri *et al.*, "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: International Committee for Computational Linguistics, Sep. 2020, pp. 1425–1447. doi: 10.18653/v1/2020.semeval-1.188.

[5] P. Alonso, R. Saini, and G. Kovacs, "TheNorth at SemEval-2020 Task 12: Hate Speech Detection Using RoBERTa," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: International Committee for Computational Linguistics, 2020, pp. 2197–2202. doi: 10.18653/v1/2020.semeval-1.292.

[6] B. Ray and A. Garain, "JU at HASOC 2020: Deep Learning with RoBERTa and Random Forest for Hate Speech and Offensive Content Identification in Indo-European Languages," 2020. [Online]. Available: http://ceur-ws.org
[7] 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, 2020, pp. 1644–1650. doi: 10.18653/v1/2020.findings-emnlp.148.
[8] G. Kaplun, A. Gurevich, T. Swisa, M. David, S. Shalev-Shwartz, and E. Malach, "Less is More: Selective Layer Finetuning with SubTuning," Jul. 2023. [Online]. Available: http://arxiv.org/abs/2302.06354
[9] M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, and Y. Elazar, "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation," May 2023. [Online]. Available: http://arxiv.org/abs/2305.16938
[10] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Federated Freeze BERT for text classification," *J. Big Data*, vol. 11, no. 1, p. 28, Feb. 2024, doi: 10.1186/s40537-024-00885-x.
[11] 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.
[12] Y. Zhang, V. Hangya, and A. Fraser, "A Study of the Class Imbalance Problem in Abusive Language Detection," in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 38–51. doi: 10.18653/v1/2024.woah-1.4.
[13] 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.
[14] J. A. Ortiz-Zambrano, C. H. Espín-Riofrío, and A. Montejo-Ráez, "Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction," *Neural Comput. Appl.*, vol. 37, no. 3, pp. 1171–1187, Jan. 2025, doi: 10.1007/s00521-024-10662-9.
[15] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, vol. 8, no. 6, pp. 4663–4678, Dec. 2022, doi: 10.1007/s40747-021-00608-2.
[16] S. Modha *et al.*, "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech," in *Forum for Information Retrieval Evaluation*, New York, NY, USA: ACM, Dec. 2021, pp. 1–3. doi: 10.1145/3503162.3503176.
[17] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *International Journal of Advances in Intelligent Informatics*, vol. 6, no. 3, p. 303, Nov. 2020, doi: 10.26555/ijain.v6i3.471.
[18] B. W. Lee, Y. S. Jang, and J. Lee, "Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 10669–10686. doi: 10.18653/v1/2021.emnlp-main.834.
[19] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.
[20] T. Mandla *et al.*, "Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages," Aug. 2021, doi: 10.1007/978-3-030-73696-5_6.
[21] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.
[22] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.
[23] CardiffNLP, "cardiffnlp/twitter-roberta-base-offensive," CardiffNLP. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive
[24] 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, 2020, pp. 1644–1650. doi: 10.18653/v1/2020.findings-emnlp.148.
[25] K. Gémes, Á. Kovács, M. Reichel, and G. Recski, "Offensive text detection on English Twitter with deep learning models and rule-based systems," 2021. [Online]. Available: https://github.com/GKingA/tuw-inf-hasoc2021.
[26] H. Ma'rifah, A. P. Wibawa, and M. I. Akbar, "Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing," *Sains, Aplikasi, Komputasi dan Teknologi Informasi*, vol. 2, no. 2, p. 70, Apr. 2020, doi: 10.30872/jsakti.v2i2.2681.
[27] Arif Bijaksana Putra Negara, "The Influence Of Applying Stopword Removal And Smote On Indonesian Sentiment Classification," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 03, pp. 172–185, Oct. 2025, doi: 10.24843/LKJITI.2023.v14.i03.p05.
[28] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692
[29] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," *JUITA : Jurnal Informatika*, vol. 10, no. 2, p. 225, Nov. 2022, doi: 10.30595/juita.v10i2.13262.
[30] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 6, no. 2, p. 129, Dec. 2021, doi: 10.30998/string.v6i2.10133.
[31] R. M. O. Cruz, W. V. de Sousa, and G. D. C. Cavalcanti, "Selecting and combining complementary feature representations and classifiers for hate speech detection," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.06721
[32] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, p. 21631, Jul. 2025, doi: 10.1038/s41598-025-05791-7.
[33] A. Demircioğlu, "Applying oversampling before cross-validation will lead to high bias in radiomics," *Sci. Rep.*, vol. 14, no. 1, p. 11563, May 2024, doi: 10.1038/s41598-024-62585-z.

[34] A. Glazkova, "A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification," Aug. 2020. [Online]. Available: http://arxiv.org/abs/2008.04636
[35] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," *Pattern Recognit.*, vol. 124, p. 108511, Apr. 2022, doi: 10.1016/j.patcog.2021.108511.
[36] R. Li, M. Liu, D. Xu, J. Gao, F. Wu, and L. Zhu, "A Review of Machine Learning Algorithms for Text Classification," 2022, pp. 226–234. doi: 10.1007/978-981-16-9229-1_14.
[37] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1301.
[38] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: 10.3390/s23042333.
[39] I. A. Rahma and L. H. Suadaa, "Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 6, pp. 1329–1340, Dec. 2023, doi: 10.25126/jtiik.2023107325.
[40] Febiana Anistya and Erwin Budi Setiawan, "Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1044–1051, Dec. 2021, doi: 10.29207/resti.v5i6.3521.
[41] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," 2023. [Online]. Available: http://gcancer.org/pdr
[42] S. Saseendran and S. Giri, "Classification of Hate Speech and Offensive Content using an approach based on DistilBERT," 2021. [Online]. Available: https://github.com/sharan0276