 ISSN:2302-8149
e-ISSN:2540-9719

Meningkatkan Deteksi Ujaran Kebencian dan Bahasa Menyinggung menggunakan *CatBoost* dengan *Embedding* Kontekstual berbasis *RoBERTa*

Enhancing Hate Speech and Offensive Language Detection using CatBoost with RoBERTa-based Contextual Embeddings

[1]Muhammad Elfarizi*, [2]Surya Agustian, [3]Fitra Kurnia, [4]Suwanto Sanjaya, [5]Fitri Insani
[1,2,3,4,5]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
[1,2,3,4,5]Jl. H.R. Soebrantas KM. 15 No. 155 Tuah Madani Kec. Tuah Madani, Pekanbaru
*e-mail: 12050113028@students.uin-suska.ac.id, surya.agustian@uin-suska.ac.id, fitra.k@uin-suska.ac.id, suwantosanjaya@uin-suska.ac.id, fitri.insani@uin-suska.ac.id

(***received*:** 20 June 2026, ***revised*:** 28 June 2026, ***accepted*:** 29 June 2026)

Abstrak

Penyebaran ujaran kebencian dan konten ofensif di platform media sosial telah menjadi masalah sosial yang sangat serius sehingga dibutuhkan sistem deteksi otomatis yang andal. Penelitian ini mengusulkan pendekatan hibrida yang memanfaatkan *frozen embedding* dari model bahasa pra-latih *cardiffnlp/twitter-roberta-base-offensive* sebagai pengekstraksi fitur semantik tingkat tinggi, dikombinasikan dengan algoritma gradient boosting CatBoost sebagai pengklasifikasi akhir. Metode dievaluasi pada dataset HASOC 2021 berbahasa Inggris melalui enam skenario eksperimen dan dibandingkan terhadap *baseline* TF-IDF menggunakan parameter bawaan CatBoost. Hasil eksperimen membuktikan pendekatan yang diusulkan mencapai Macro *F1-Score* 0.7924 pada klasifikasi biner (Tugas-1A) dan 0.6113 pada klasifikasi multikelas (Tugas-1B), melampaui *baseline* TF-IDF (0.7724 dan 0.5798). Sistem yang dihasilkan terdapat peningkatan dan mampu masuk dalam tim-tim terbaik pada papan peringkat resmi HASOC 2021 tanpa memerlukan proses *fine-tuning* yang mahal secara komputasi.

Kata kunci: CatBoost, deteksi ujaran kebencian, HASOC 2021, RoBERTa Embedding, TF-IDF

Abstract

The widespread dissemination of hate speech and offensive content on social media platforms has become a critical societal issue, highlighting the need for reliable automated detection systems. This study proposes a hybrid approach that leverages frozen embeddings from the pre-trained language model cardiffnlp/twitter-roberta-base-offensive as a high-level semantic feature extractor, combined with the CatBoost gradient boosting algorithm as the final classifier. The proposed method was evaluated on the HASOC 2021 English dataset through six experimental scenarios and compared with a TF-IDF baseline using CatBoost's default hyperparameters. Experimental results demonstrate that the proposed approach achieved a Macro F1-score of 0.7924 for the binary classification task (Task 1A) and 0.6113 for the multiclass classification task (Task 1B), outperforming the TF-IDF baseline, which achieved scores of 0.7724 and 0.5798, respectively. The proposed system demonstrated a clear performance improvement and achieved results comparable to those of the top-ranked teams on the official HASOC 2021 leaderboard, while avoiding the computational cost associated with fine-tuning large pre-trained language models.

Keywords: *CatBoost, HASOC 2021, hate speech detection, RoBERTa Embedding, TF-IDF.*

1 Pendahuluan

Maraknya ujaran kebencian dan konten yang menyinggung di platform media sosial telah berkembang menjadi masalah sosial yang sangat mengkhawatirkan seiring dengan pesatnya perkembangan platform digital dan peningkatan jumlah pengguna internet secara global [1]. Ujaran

 ISSN:2302-8149
e-ISSN:2540-9719

kebencian secara umum didefinisikan sebagai setiap ungkapan, tulisan, atau tindakan provokatif yang secara sengaja menyerang, merendahkan, atau mempermalukan individu atau kelompok berdasarkan atribut identitas tertentu seperti ras, agama, gender, atau etnis. Paparan yang terus-menerus terhadap ujaran kebencian di ruang digital telah terbukti secara empiris menimbulkan dampak psikologis yang serius pada korban yang menjadi sasaran, mulai dari perasaan dikucilkan secara sosial dan depresi hingga risiko bunuh diri [2]. Mengingat volume yang sangat besar dan penyebaran konten semacam itu yang tidak terkendali, moderasi manual tidak lagi efisien, sehingga sistem deteksi otomatis secara real-time menjadi kebutuhan yang mendesak.

Tantangan utama dalam membangun sistem pendeteksi ujaran kebencian terletak pada tingginya tingkat ambiguitas dan kompleksitas linguistik yang terdapat dalam teks media sosial yang dibuat oleh pengguna. Postingan sering kali tidak terstruktur, sarat dengan bahasa gaul, singkatan non-standar, dan peralihan kode bahasa antar berbagai bahasa [3]. Selain itu, batas antara ujaran kebencian dan bahasa yang sekadar menyinggung atau kasar sangatlah subjektif dan memerlukan pemahaman kontekstual yang mendalam. Ketidakseimbangan kelas yang parah dalam kategori sentimen negatif juga menjadi tantangan besar bagi model klasifikasi [4]. Untuk memajukan penelitian, komunitas ilmiah secara rutin menyelenggarakan kompetisi benchmark, terutama tugas bersama Hate Speech and Offensive Content Identification (HASOC), yang menyediakan dataset dan protokol evaluasi terstandarisasi.

Secara historis, klasifikasi teks bergantung pada metode ekstraksi fitur statistik seperti TF-IDF (Term Frequency-Inverse Document Frequency). Namun, literatur terkini secara konsisten menunjukkan bahwa metode konvensional ini sering kali gagal menangkap makna semantik laten dan konteks struktural antar kata, karena metode tersebut beroperasi murni berdasarkan statistik frekuensi kata tanpa memperhitungkan posisi kata atau konteks relasional [1]. Akibatnya, model TF-IDF tidak dapat membedakan kalimat yang memiliki distribusi kata pada tingkat permukaan yang hampir identik tetapi memiliki makna yang bertolak belakang tergantung pada konteksnya.

Sebagai solusi atas keterbatasan metode statistik klasik, arsitektur berbasis Transformer seperti BERT (Bidirectional Encoder Representations from Transformers) dan RoBERTa (Robustly Optimized BERT Pretraining Approach) telah menjadi standar mutakhir yang mendominasi sebagian besar tolok ukur Pemrosesan Bahasa Alami (NLP) [5]. Secara khusus, varian model bahasa yang telah dilatih sebelumnya, yaitu *cardiffnlp/twitter-roberta-base-offensive*, yang dilatih secara khusus pada ratusan juta postingan Twitter berbahasa Inggris, telah terbukti menghasilkan representasi vektor yang lebih kaya secara kontekstual yang jauh lebih baik dalam menangkap nuansa bahasa media sosial informal [6].

Meskipun keunggulan model *Transformer* telah terbukti, proses *fine-tuning* penuh membutuhkan sumber daya komputasi yang sangat besar. Sebagai alternatif yang lebih efisien, penelitian terbaru mengusulkan pendekatan hibrida: menggunakan Transformer semata-mata sebagai ekstraktor fitur statis yang dibekukan, kemudian memasukkan *embedding* yang dihasilkan ke dalam algoritma Gradient Boosting sebagai klasifikasi hilir [7]. Di antara keluarga boosting, CatBoost telah menunjukkan kinerja prediktif yang tinggi secara konsisten sekaligus tetap tangguh terhadap overfitting pada tugas pemodelan teks [8].

Dengan latar belakang tersebut, penelitian ini secara empiris membandingkan representasi fitur semantik dari *embedding cardiffnlp/twitter-roberta-base-offensive* yang dibekukan dengan fitur TF-IDF klasik untuk mendeteksi ujaran kebencian. Untuk memastikan objektivitas, CatBoost diterapkan secara konsisten dengan konfigurasi *default*nya dan tanpa penyesuaian hiperparameter pada kedua alur kerja, sehingga perbedaan kinerja dapat dikaitkan semata-mata dengan kualitas representasi fitur. Evaluasi dilakukan pada dataset benchmark HASOC 2021 bahasa Inggris untuk dua tugas: klasifikasi biner (Tugas-1A) dan klasifikasi multi-kelas (Tugas-1B).

2 Tinjauan Literatur

Dalam kajian analisis sentimen dan klasifikasi teks, ekstraksi fitur merupakan tahapan krusial untuk mengubah data teks yang tidak terstruktur menjadi representasi numerik. Sebagai salah satu pendekatan statistik konvensional, metode *Term Frequency-Inverse Document Frequency* (TF-IDF) sangat umum digunakan karena kemampuannya dalam menilai relevansi suatu kata dengan

 ISSN:2302-8149
e-ISSN:2540-9719

mempertimbangkan frekuensi kemunculannya di dalam sebuah dokumen dan seberapa jarang kata tersebut ditemukan di seluruh kumpulan dokumen [9]. Meskipun TF-IDF secara efektif mampu menyoroti istilah penting, metode ini memiliki keterbatasan mendasar karena gagal mewakili informasi semantik dan hubungan kontekstual antar kata terutama yang berhubungan dengan ujaran kebencian dan kalimat menyinggung[10].

Penelitian mengenai deteksi ujaran kebencian telah berkembang pesat untuk mengatasi keterbatasan tersebut, terutama melalui seri benchmark HASOC. Mandl et al, (2021) merilis dataset resmi HASOC 2021 beserta laporan kompetisi yang komprehensif[3]. Hasil kompetisi tersebut menunjukkan dominasi arsitektur *Transformer* yang sangat jelas di papan peringkat, meskipun pembedaan nuansa linguistik yang lebih halus antara kelas Ujaran Kebencian (*Hate*), Konten Ofensif (*Offensive*), dan Umpatan (*Profane*)[3].

Essa et al, (2023) mengusulkan model hibrida yang menggabungkan representasi semantik berbasis BERT dengan LightGBM sebagai klasifikasi hilir untuk mendeteksi berita palsu. Penelitian mereka menunjukkan bahwa metode hibrida tersebut lebih unggul dibandingkan penggunaan BERT atau LightGBM secara terpisah, sehingga membuktikan bahwa penggunaan Transformer sebagai ekstraktor fitur yang dipadukan dengan algoritma ML konvensional merupakan pendekatan yang valid, efisien secara komputasi, dan kompetitif[7].

Ahn et al, (2023) membuktikan keunggulan pendekatan ensembel *Gradient Boosting* yang menggabungkan XGBoost, LightGBM, dan CatBoost. Hasil penelitian mereka menunjukkan bahwa ensembel dari ketiga varian boosting tersebut secara konsisten menghasilkan akurasi prediksi yang lebih tinggi dan lebih stabil dibandingkan dengan model individu mana pun, sehingga memberikan dasar yang kuat untuk memilih CatBoost sebagai klasifikasi dalam penelitian ini[11].

Glazkova et al, (2021) menyempurnakan model RoBERTa khusus untuk Twitter dengan domain untuk HASOC 2021 dan berhasil meraih peringkat ketiga pada Tugas 1A bahasa Inggris dengan skor *F1* sebesar 0,8199. Penelitian mereka membuktikan bahwa model yang dilatih terlebih dahulu secara khusus pada korpus Twitter jauh lebih efektif dalam mendeteksi konten media sosial yang negatif, yang menjadi dasar langsung dalam pemilihan model *cardiffnlp/twitter-roberta-base-offensive* dalam studi ini[12].

Putri & Ardiansyah, (2023) membandingkan BERT dan RoBERTa untuk analisis sentimen bahasa Indonesia dan menemukan bahwa RoBERTa menghasilkan representasi fitur yang lebih baik serta skor *F1* yang lebih tinggi, hal ini dikaitkan dengan strategi pra-pelatihan yang lebih unggul, terutama penghapusan tujuan Prediksi Kalimat Berikutnya (NSP) dan penggunaan masking dinamis. Kumar et al., (2021) menggunakan gabungan algoritma ML klasik (SVM, RF, LR) untuk HASOC, dan menyimpulkan bahwa metode-metode ini masih tertinggal dibandingkan Deep Learning dalam menangkap konteks semantik yang kompleks[13].

Herwanto et al, (2021) menunjukkan bahwa *embedding* kontekstual (BERT/RoBERTa) secara signifikan mengurangi kesalahan klasifikasi yang disebabkan oleh sarkasme dan ambiguitas linguistik dibandingkan dengan *embedding* kata statis seperti *Word2Vec* dan *FastText*[15]. Alkomah & Ma, (2022) mengamati dalam survei mereka bahwa tren penelitian saat ini mengarah pada Pembelajaran Mendalam Hibrida, di mana model Transformer menangani pemahaman teks dan algoritma lain mengelola klasifikasi pada data yang tidak seimbang[1]. Liu et al, (2019) menunjukkan bahwa penggabungan fitur BERTweet dengan fitur statistik pelengkap meningkatkan ketahanan model terhadap variasi bahasa di Twitter[2], sementara Guo et al, (2023) memvalidasi keunggulan CatBoost dalam menangani data kompleks dan mencegah *overfitting* lebih baik daripada algoritma *boosting* lainnya[8].

Berbeda dengan penelitian-penelitian terdahulu yang umumnya mengandalkan proses *full fine-tuning* pada arsitektur *Transformer* dengan biaya komputasi yang sangat mahal seperti yang dilakukan oleh sebagian besar tim pemuncak klasemen kompetisi HASOC 2021, penelitian ini menawarkan kebaruan berupa pendekatan hibrida yang sangat efisien secara komputasi. Kebaruan utama penelitian ini terletak pada pemanfaatan model bahasa spesifik domain *cardiffnlp/twitter-roberta-base-offensive* secara eksklusif sebagai pengekstraksi fitur statis (*frozen embedding*), yang kemudian diklasifikasikan menggunakan algoritma CatBoost dengan konfigurasi parameter bawaannya (*default*). Pendekatan tanpa *fine-tuning* dan tanpa optimasi hyperparameter ini secara sengaja dirancang untuk membuktikan bahwa kualitas representasi semantik laten dari frozen *embedding* RoBERTa sudah cukup tangguh untuk menghasilkan kinerja *state-of-the-art*. Selain itu, penelitian ini juga memberikan kebaruan

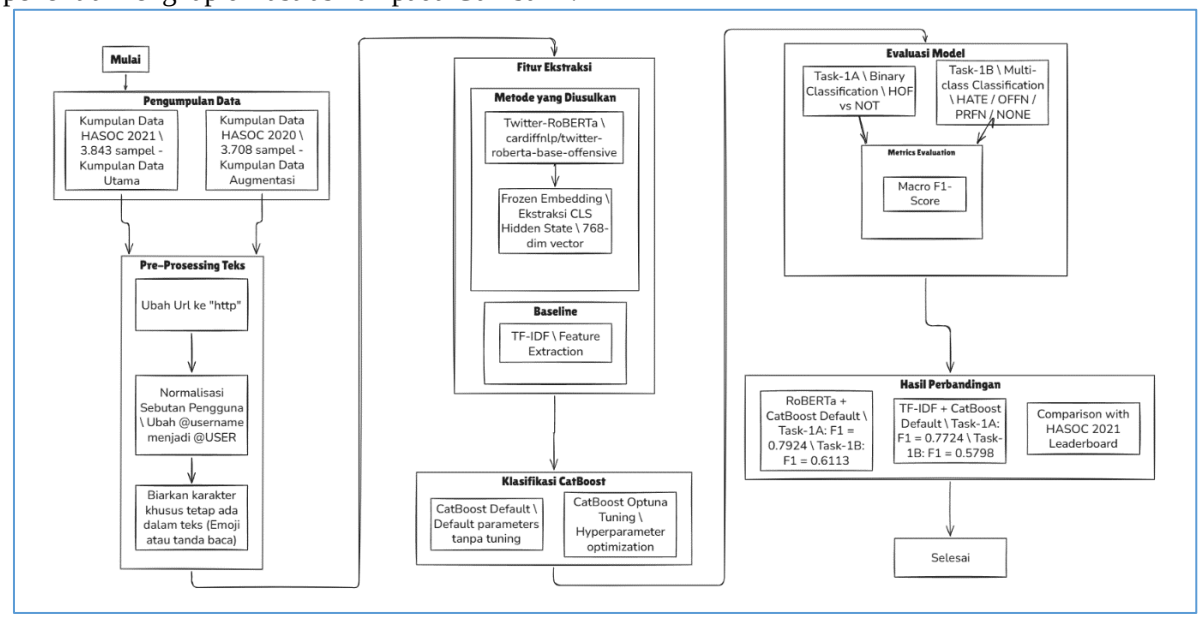
 ISSN:2302-8149
e-ISSN:2540-9719

analitis dengan menyoroti perbedaan perilaku yang sangat kontras antara metode statistik (TF-IDF) dan representasi kontekstual (RoBERTa) saat dihadapkan pada augmentasi data lintas tahun.

3 Metode Penelitian

3.1 Alur Penelitian

Penelitian ini dilakukan melalui alur kerja sistematis yang dimulai dari pengumpulan data dan diakhiri dengan evaluasi komparatif terhadap model klasifikasi akhir. Teks yang dikumpulkan melalui tahap prapemrosesan untuk menghilangkan noise sebelum diekstraksi menjadi dua representasi fitur yang berbeda: (1) fitur statistik TF-IDF sebagai *baseline*, dan (2) fitur *deep learning* tingkat tinggi yang menggunakan *embedding cardiffnlp/twitter-roberta-base-offensive* yang dibekukan sebagai metode yang diusulkan [7]. Fitur dari masing-masing metode kemudian diklasifikasikan secara eksklusif menggunakan algoritma CatBoost dengan konfigurasi *default*-nya, dan dievaluasi menggunakan *Macro F1-Score* untuk memastikan perbandingan yang seimbang dan objektif. Alur penelitian lengkap diilustrasikan pada Gambar 1.



Gambar 1 Alur penelitian

3.2 Dataset

Penelitian ini menggunakan dua kumpulan data dari tugas bersama HASOC *(Hate Speech and Offensive Content Identification)*, yang diselenggarakan sebagai bagian dari rangkaian konferensi FIRE *(Forum for Information Retrieval Evaluation)*. Kedua kumpulan data tersebut dipilih karena berfungsi sebagai tolok ukur yang diakui secara internasional di kalangan komunitas penelitian NLP untuk tugas identifikasi ujaran kebencian dalam bahasa Inggris.

a. Dataset HASOC 2020
   Kumpulan data HASOC 2020 dalam bahasa Inggris digunakan secara eksklusif sebagai data tambahan untuk memperkaya keragaman pola linguistik selama proses pelatihan model [4]. Kumpulan data ini berisi 3.708 sampel teks yang dibagi ke dalam dua tugas klasifikasi. Distribusi label kumpulan data HASOC 2020 disajikan pada Tabel 1.

Tabel 1 Distribusi label pada kumpulan data HASOC 2020 (data augmentasi)

Tugas	Label	Deskripsi	Sampel
Subtask 1A	HOF	Hate and Offensive	1,856
Subtask 1A	NOT	Non Hate Offensive	1,852
Subtask 1B	NONE	Non-Hate	1,852
Subtask 1B	PRFN	Profane	1,377

 ISSN:2302-8149
e-ISSN:2540-9719

Subtask 1B	HATE	Hate Speech	158
Subtask 1B	OFFN	Offensive	321
Total	—	—	3,708

b. Dataset HASOC 2021

Kumpulan data utama dalam penelitian ini bersumber dari tugas bersama HASOC 2021 dalam bahasa Inggris, yang berfokus pada identifikasi konten yang menyinggung dan ujaran kebencian [2]. Bagian pelatihan berisi 3.843 sampel dan bagian pengujian berisi 1.281 sampel, yang semuanya dianotasi secara manual oleh tim anotasi khusus [12]. Dengan menggabungkan bagian pelatihan HASOC 2020 dan 2021, diperoleh total 7.551 sampel pelatihan, yang memberikan generalisasi linguistik yang lebih luas [16]. Distribusi label dari kumpulan data utama HASOC 2021 ditunjukkan pada Tabel 2.

Tabel 2. Distribusi label pada kumpulan data HASOC 2021 (kumpulan data utama)

Tugas	Label	Deskripsi	Sampel
Subtask 1A	HOF	Hate and Offensive	2,501
Subtask 1A	NOT	Non Hate Offensive	1,342
Subtask 1B	NONE	Non-Hate	1,342
Subtask 1B	PRFN	Profane	1,196
Subtask 1B	HATE	Hate Speech	683
Subtask 1B	OFFN	Offensive	622
Total	—	—	3,843

c. Pembagian Data Validasi

Untuk memastikan evaluasi model yang objektif sebelum dihadapkan pada data uji resmi (*blind test set*), kumpulan data latih yang telah dikumpulkan tidak digunakan seluruhnya untuk proses pembelajaran model. Data tersebut dibagi menjadi dua subset utama, yaitu data latih (training set) dan data validasi (validation set). Pembagian data validasi ini berfungsi sebagai tolok ukur internal untuk memantau kinerja algoritma CatBoost selama fase pelatihan, mencegah terjadinya *overfitting*, serta menyeleksi konfigurasi metode ekstraksi fitur yang paling optimal. Secara spesifik, dari total keseluruhan sampel pelatihan pada masing-masing skenario, sebanyak 80% data dialokasikan sebagai data latih dan 20% sisanya dialokasikan secara acak sebagai data validasi. Pemisahan ini memastikan bahwa performa model yang dilaporkan pada fase validasi merupakan hasil prediksi pada data yang benar-benar belum pernah dilihat oleh model selama proses pelatihan.

3.3 Pra-pemrosesan Teks

Teks media sosial bersifat sangat dinamis, tidak terstruktur, dan berisik, sehingga memerlukan langkah-langkah normalisasi untuk membersihkan data masukan sebelum diproses lebih lanjut. Namun, dalam penelitian ini, proses *Case Folding* (mengubah teks menjadi huruf kecil) sengaja diabaikan, dan karakter khusus serta tanda baca dibiarkan utuh tanpa dihapus. Keputusan ini diambil untuk mempertahankan koherensi sintaksis dan semantik asli dari teks media sosial [17]. Penggunaan huruf kapital (yang sering kali menunjukkan penekanan emosional atau kemarahan) dan keberadaan karakter khusus sering kali membawa sinyal laten yang krun sial untuk mendeteksi intensitas ujaran kebencian. Sinyal-sinyal ini dapat dipetakan dengan sangat efektif oleh *tokenizer* sub-kata dari arsitektur RoBERTa.

Oleh karena itu, alur pemrosesan awal yang diterapkan dalam penelitian ini hanya terbatas pada pembersihan teknis, yang terdiri dari langkah-langkah berikut:

a. Normalisasi URL

Mengganti semua tautan *web* dengan token netral umum '*http*'. Meskipun tautan yang sebenarnya tidak mengandung informasi semantik yang relevan, menormalisasikannya

 ISSN:2302-8149
e-ISSN:2540-9719

menjadi token standar membantu model mengenali adanya tautan tanpa menambah kebisingan kosakata [6].

b. Normalisasi Penyebutan Pengguna
Mengganti semua tag *@username* dengan token netral umum '*@USER*', sesuai dengan standar normalisasi teks Twitter [6].

c. Membiarkan karakter khusus
Pada bagian karakter khusus seperti tanda baca atau emoji tidak dilakukan perubahan karena model dapat mendeteksi karakter tersebut yang nantinya akan diubah ke dalam token oleh RoBERTa[2].

3.4 TF-IDF *(Feature Extraction Baseline)*

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode ekstraksi fitur statistik tradisional yang mengubah data teks menjadi matriks numerik dengan memberikan bobot pada frekuensi suatu istilah dalam suatu dokumen relatif terhadap tingkat keunikannya di seluruh korpus [18]. TF-IDF efektif dalam menyoroti kosakata yang spesifik pada korpus, tetapi pada dasarnya memiliki keterbatasan karena ketidakmampuannya menangkap makna semantik laten dan hubungan kontekstual antar kata [7]. Pendekatan ini digunakan semata-mata sebagai acuan dasar untuk mengukur batas atas kinerja metode statistik konvensional. Secara matematis, TF-IDF dihitung melalui tiga persamaan.

$$TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \quad (1)$$

Dimana

$TF(t,d)$ = Nilai frekuensi kata t dalam dokumen d
$n(t,d)$ = jumlah kemunculan istilah t dalam dokumen d
$\sum_k n_{k,d}$ = jumlah total semua istilah dalam dokumen d

$$IDF(t,D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad (2)$$

Dimana

$IDF(t,D)$ = Nilai kepentingan kata di seluruh koleksi dokumen
N = Jumlah total dokumen dalam korpus D
$|\{d \in D : t \in d\}|$ = Jumlah dokumen yang mengandung istilah t setidaknya sekali

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (3)$$

Keterangan

t = Kata atau *term* spesifik yang sedang dihitung bobotnya
d = Dokumen atau baris teks spesifik tempat kata tersebut berada
D = *Korpus* atau keseluruhan koleksi dokumen yang digunakan dalam penelitian
$n_{t,d}$ = Jumlah kemunculan kata t dalam dokumen d
$\sum_k n_{k,d}$ = Total seluruh kata yang terdapat dalam dokumen d
N = Total jumlah dokumen yang terdapat dalam keseluruhan korpus D
$|\{d \in D : t \in d\}|$ = Jumlah dokumen yang mengandung kata t setidaknya satu kali
log = Fungsi logaritma yang digunakan untuk meredam nilai frekuensi agar Bobot yang dihasilkan tetap proporsional dan tidak didominasi oleh

 ISSN:2302-8149
e-ISSN:2540-9719

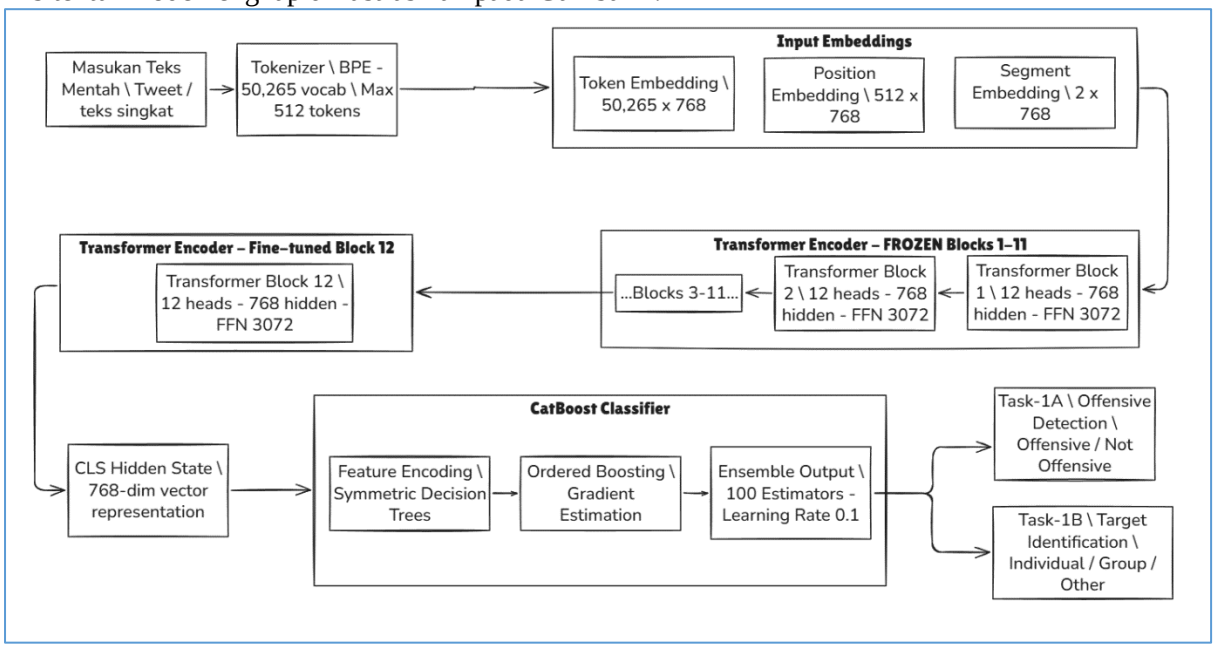
jumlah dokumen yang sangat besar

3.5 RoBERTa *(Proposed Method)*

RoBERTa *(Robustly Optimized BERT Pretraining Approach)* adalah varian BERT yang telah dioptimalkan dengan modifikasi penting pada strategi pra-pelatihan [2]. Keunggulannya dibandingkan BERT meliputi penghapusan tujuan *Next Sentence Prediction* (NSP), penerapan *dynamic masking* yang memastikan setiap token disembunyikan secara berbeda di setiap epoch pelatihan, serta pra-pelatihan pada dataset yang jauh lebih besar selama periode yang lebih lama [13].

Penelitian ini menggunakan varian model *cardiffnlp/twitter-roberta-base-offensive*, yang secara khusus telah dilatih sebelumnya menggunakan ratusan juta cuitan Twitter berbahasa Inggris. Pemilihan varian khusus Twitter ini didasarkan pada temuan Glazkova et al., (2021), yang menunjukkan bahwa pelatihan awal khusus domain pada korpus Twitter menghasilkan representasi fitur yang jauh lebih relevan untuk mendeteksi konten media sosial yang bernada negatif.

Ekstraksi fitur dilakukan melalui skema *embedding* beku representasi *vektor* dari keadaan tersembunyi token [CLS] pada lapisan model terakhir diekstraksi secara langsung, tanpa melakukan penyempurnaan terhadap parameter jaringan internal apa pun. Token [CLS] dipilih karena secara arsitektural dirancang untuk menggabungkan representasi semantik dari seluruh kalimat masukan menjadi satu vektor padat berdimensi 768. Dengan pendekatan ini, model RoBERTa berfungsi secara eksklusif sebagai ekstraktor fitur semantik statis yang kaya konteks namun sangat efisien secara komputasi, tidak memerlukan perangkat keras GPU khusus maupun waktu pelatihan yang lama [7]. Arsitektur model lengkap diilustrasikan pada Gambar 2.



Gambar 2. Arsitektur *cardiffnlp/twitter-roberta-base-offensive* + CatBoost

3.6 CatBoost *(Classifier Algorithm)*

CatBoost *(Categorical Boosting)* adalah kerangka kerja pembelajaran mesin *ensembel* mutakhir yang dibangun berdasarkan prinsip *Gradient Boosting Decision Tree* (GBDT), yang menggabungkan prediksi dari serangkaian pohon keputusan secara berurutan [8]. Berbeda dengan algoritma *boosting* konvensional seperti XGBoost dan LightGBM, CatBoost dirancang berdasarkan pohon keputusan simetris yang mengurangi derajat kebebasan model, secara signifikan meningkatkan kecepatan komputasi, dan menawarkan ketahanan yang lebih baik terhadap *overfitting* [11].

Karena penelitian ini mengevaluasi dua tugas dengan jumlah target yang berbeda, fungsi kerugian (*loss function*) yang dioptimalkan oleh CatBoost disesuaikan dengan jenis tugas klasifikasi tersebut. Untuk klasifikasi biner pada Tugas-1A [1], CatBoost menggunakan fungsi *Binary Cross Entropy* (Log Loss), yang dirumuskan sebagai berikut.

 ISSN:2302-8149
e-ISSN:2540-9719

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (4)$$

Dimana

L	=	Nilai *loss* atau kerugian yang dihasilkan
y_i	=	Label aktual/target untuk data ke-i
p_i	=	Probabilitas hasil prediksi model bahwa data ke-i termasuk dalam kelas positif
N	=	Total jumlah sampel data dalam proses evaluasi

Dalam semua eksperimen, CatBoost dijalankan secara konsisten dengan konfigurasi *default*nya tanpa optimasi *hiperparameter*. Hal ini dilakukan secara sengaja untuk memastikan bahwa setiap perbedaan kinerja yang terukur antara *pipeline* TF-IDF dan RoBERTa semata-mata disebabkan oleh kualitas metode representasi fitur, bukan oleh penyetelan klasifikasi [8].

Sedangkan untuk klasifikasi multikelas pada Tugas-1B (4 kelas target), fungsi kerugian yang dioptimalkan adalah *Multiclass Cross Entropy* (*Categorical Cross Entropy*), yang dirumuskan sebagai berikut.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c} \log(p_{i,c}) \quad (5)$$

Dimana

M	=	Jumlah total kelas target (pada Tugas-1B)
$y_{i,c}$	=	ndikator biner (bernilai 1 jika kelas *c* adalah label yang benar untuk data ke-i, dan 0 jika salah)
$p_{i,c}$	=	Probabilitas hasil prediksi model bahwa data ke-i termasuk dalam kelas c

3.7 Matriks Evaluasi

Tahap terakhir dari kerangka kerja penelitian ini adalah pengujian dan evaluasi kinerja model klasifikasi. Mengacu pada pedoman evaluasi resmi yang ditetapkan pada kompetisi FIRE HASOC 2021[3], metrik utama yang digunakan untuk mengukur dan membandingkan performa antar model adalah *Macro-Averaged F1-Score* [18].

Penggunaan metrik *Macro* secara khusus dipilih karena metrik ini sangat tangguh dan objektif dalam menangani masalah ketidakseimbangan distribusi kelas (*class imbalance*) yang umum terjadi pada data media sosial. Pendekatan ini mengevaluasi kinerja model dengan memperlakukan kelas minoritas dan mayoritas secara setara tanpa memandang bias jumlah sampel pada masing-masing kelas. Secara matematis, sebelum mendapatkan agregasi nilai *Macro*, sistem akan terlebih dahulu menghitung nilai dasar dari *F1-Score* pada setiap kelas secara independen. *F1-Score* itu sendiri merupakan rata-rata harmonik dari tingkat presisi (*precision*) dan perolehan (*recall*), yang dirumuskan pada Persamaan 6.

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

F1-Score merupakan rata-rata harmonik dari *Precision* dan *Recall*. Metrik ini dipilih sebagai metrik evaluasi utama karena merupakan ukuran yang paling adil dan tangguh dalam menangani ketidakseimbangan distribusi kelas, yang sering ditemui dalam dataset media sosial. Metrik yang

 ISSN:2302-8149
e-ISSN:2540-9719

dilaporkan adalah *Macro-Averaged F1-Score*, yang memperlakukan semua kelas secara setara tanpa memandang frekuensinya.

Setelah nilai *F1-Score* dari masing-masing kelas target didapatkan, langkah selanjutnya adalah menghitung *Macro-Averaged F1-Score* dengan cara mencari nilai rata-rata aritmatika dari seluruh kelas tersebut, seperti yang dirumuskan pada Persamaan 8.

$$Macro\ F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i \quad (7)$$

Keterangan

c = Total jumlah kelas target (misalnya 2 untuk Tugas-1A, dan 4 untuk Tugas-1B)

$F1_i$ = Nilai *F1-Score* untuk kelas ke-i

4 Hasil dan Pembahasan

4.1 Hasil Klasifikasi Biner (Tugas-1A)

Evaluasi pertama berfokus pada Tugas-1A: klasifikasi *biner* yang membedakan antara teks yang mengandung ujaran kebencian dan menyinggung (HOF) dengan teks yang menyinggung namun tidak mengandung ujaran kebencian (NOT). Eksperimen dilakukan pada enam skenario berbeda yang mencakup variasi dalam data pelatihan (hanya HASOC 2021 vs. gabungan HASOC 2020+2021), metode ekstraksi fitur (TF-IDF vs. RoBERTa), dan konfigurasi klasifikasi (*default* vs. penyetelan *hiperparameter Optuna*). Hasilnya dirangkum dalam Tabel 3.

Tabel 3 Hasil evaluasi untuk tugas-1A pada data validasi

Skenario	Data Training	Ekstaraksi Fitur	Konfigurasi CatBoost	Macro Avg F1 Score
1	Hasoc 2021	TF-IDF	*Default*	0.7457
2	Hasoc 2020 + 2021	TF-IDF	*Default*	0.7457
3	Hasoc 2021	RoBERTa Embedding	*Default*	0.7964
4 (Terbaik)	Hasoc 2020 + 2021	RoBERTa Embedding	*Default*	0.8019
5	Hasoc 2020 + 2021	TF-IDF	Optuna	0.7358
6	Hasoc 2020 + 2021	RoBERTa Embedding	Optuna	0.7813

Skenario 4, yang menggunakan fitur *embedding* RoBERTa dengan data pelatihan yang diperluas (gabungan HASOC 2020+2021), mencapai kinerja klasifikasi *biner* terbaik dengan nilai *Macro F1-Score* sebesar 0,8019. Skenario 3 (hanya HASOC 2021) memiliki peningkatan yang dibuktikan dengan skor 0,7964 dibandingkan dengan TF-IDF yang hanya memiliki nilai *Macro F1 – Score* sebesar 0,7457, yang mencerminkan kematangan linguistik dari pengetahuan yang telah dilatih sebelumnya oleh RoBERTa. Sebuah anomali yang mencolok muncul pada Skenario 5, di mana penyetelan *hiperparameter Optuna* pada *baseline* TF-IDF secara drastis menurunkan kinerja menjadi 0,7358, yang menunjukkan bahwa fitur TF-IDF sangat rentan terhadap *overfitting* dalam ruang pencarian *hiperparameter* yang kompleks.

 ISSN:2302-8149
e-ISSN:2540-9719

4.2 Hasil Klasifikasi Multi-Kelas (Tugas-1B)

Evaluasi kedua ini menargetkan Tugas-1B: klasifikasi multi-kelas tingkat detail yang mengelompokkan sampel ke dalam empat kategori yaitu, Ucapan Kebencian (HATE), Menyinggung (OFFN), Kata-kata Kasar (PRFN), dan Bukan Ucapan Kebencian (NONE). Enam skenario yang identik diterapkan untuk menjaga konsistensi dan keterbandingan. Hasilnya disajikan dalam Tabel 4.

Tabel 4 Hasil evaluasi untuk tugas-1B pada data validasi

Skenario	Data Training	Ekstaraksi Fitur	Konfigurasi CatBoost	Macro Avg F1 Score
1	Hasoc 2021	TF-IDF	*Default*	0.5910
2	Hasoc 2020 + 2021	TF-IDF	*Default*	0.5910
3	Hasoc 2021	RoBERTa Embedding	*Default*	0.6271
4 (Terbaik)	Hasoc 2020 + 2021	RoBERTa Embedding	*Default*	0.6271
5	Hasoc 2020 + 2021	TF-IDF	Optuna	0.5837
6	Hasoc 2020 + 2021	RoBERTa Embedding	Optuna	0.6144

Fitur *embedding* RoBERTa (Skenario 3 dan 4) sekali lagi unggul dalam tugas multi-kelas. Penurunan kinerja secara keseluruhan dari Tugas-1A ke Tugas-1B memang sudah diperkirakan, mengingat tingginya ambiguitas linguistik dalam membedakan batas-batas yang halus antara kelas "Hate", '*Offensive*', dan "*Profane*" [3]. Keunggulan konsisten RoBERTa pada kedua tugas tersebut memperkuat validasi temuan utama studi ini.

4.3 Perbandingan Kinerja TF-IDF dengan Embedding RoBERTa

Berbagai eksperimen pada kedua tugas klasifikasi secara konsisten menunjukkan bahwa *embedding* RoBERTa yang dibekukan *(frozen)* secara signifikan lebih unggul daripada fitur TF-IDF klasik. Perbedaan Macro F1 antara kedua metode dalam kondisi yang identik (CatBoost *default*, gabungan dataset 2020+2021) mencapai ±0,0562 pada Tugas-1A (0.7457 vs. 0.8019) dan ±0,0361 pada Tugas-1B yang dimana *F1-score* terbaik diraih cukup dengan dataset 2021 saja dengan perbandingan (0,6271 vs. 0,5910). Hal ini sepenuhnya konsisten dengan landasan teoretis arsitektur *Transformer* vektor keadaan tersembunyi dapat menangkap konteks tingkat kalimat jauh lebih kaya daripada pemrosesan frekuensi kata [6].

Sebagai pendekatan berbasis frekuensi, TF-IDF efektif dalam menonjolkan kosakata yang spesifik pada *korpus*, namun pada dasarnya tidak mampu menangkap makna semantik laten dari sebuah kalimat secara keseluruhan [7]. Misalnya, dua kalimat dengan distribusi kata yang hampir identik tetapi memiliki makna yang berlawanan akan menghasilkan *vektor* TF-IDF yang sangat mirip, sementara RoBERTa dapat membedakannya melalui mekanisme perhatiannya yang memodelkan posisi kata dan hubungan semantik antar-token.

4.4 Pengujian Dengan Data Test Resmi Hasoc 2021

Guna mengukur ketangguhan dan kemampuan generalisasi model pada skenario dunia nyata, model dievaluasi lebih lanjut menggunakan kumpulan data uji (blind test set) resmi yang dirilis oleh pihak penyelenggara FIRE HASOC 2021. Data uji ini terdiri dari 1.281 cuitan Twitter berbahasa Inggris yang tidak pernah diikutsertakan dalam proses pelatihan maupun validasi (Mandl et al., 2021). Evaluasi difokuskan pada 3 skenario dengan konfigurasi ekstraksi fitur dan dataset terbaik dari fase sebelumnya

a. Hasil Pengujian Klasifikasi Biner (Tugas-1A)

 ISSN:2302-8149
e-ISSN:2540-9719

Pada Tugas-1A (membedakan konten Hate/Offensive dengan Non-Hate), fitur *embedding* kontekstual dari RoBERTa secara konsisten menunjukkan keunggulannya di atas metode statistik klasik TF-IDF saat berhadapan dengan data uji resmi. Rincian 3 hasil skenario terbaik untuk Tugas-1A disajikan pada Tabel 5.

Tabel 5 Top 3 skenario terbaik pada data uji resmi HASOC 2021 (tugas-1A)

Skenario	Data Training	Ekstaraksi Fitur	Konfigurasi CatBoost	Macro Avg F1 Score
1	Hasoc 2020	TF-IDF	*Default*	0.7460
2	Hasoc 2020 + 2021	TF-IDF	*Default*	0.7724
3 (Terbaik)	Hasoc 2020 + 2021	RoBERTa Embedding	*Default*	0.7924

b. Hasil Pengujian Klasifikasi Multikelas (Tugas-1B)
Untuk Tugas-1B yang memiliki tingkat granularitas kelas yang lebih rumit (Hate, Profane, Offensive, None), representasi semantik RoBERTa terbukti jauh lebih mampu memetakan nuansa bahasa yang ambigu dibandingkan TF-IDF. Rincian kinerja 3 skenario terbaik pada data uji Tugas-1B dapat dilihat pada Tabel 6.

Tabel 6 Top 3 skenario terbaik pada data uji resmi HASOC 2021 (tugas-1B)

Skenario	Data Training	Ekstaraksi Fitur	Konfigurasi CatBoost	Macro Avg F1 Score
1	Hasoc 2020	TF-IDF	*Default*	0.5471
2	Hasoc 2020 + 2021	TF-IDF	*Default*	0.5798
3 (Terbaik)	Hasoc 2020 + 2021	RoBERTa Embedding	*Default*	0.6113

4.5 Analisis Dampak Augmentasi Data

Penambahan dataset HASOC 2020 sebagai data augmentasi memperlihatkan fenomena evaluasi yang sangat menarik antara fase validasi dan fase pengujian (*testing*). Pada fase validasi, augmentasi data terlihat tidak memberikan dampak peningkatan apa pun pada metode TF-IDF, di mana skornya stagnan pada 0,7457 (Tugas-1A) dan 0,5910 (Tugas-1B). Namun, perbedaan perilaku yang sangat mencolok dari augmentasi ini baru terbukti ketika model dievaluasi menggunakan data uji resmi (*blind test set*) yang berisi kosakata yang belum pernah dilihat oleh model sebelumnya.

Pada hasil data *test*, augmentasi sukses membantu TF-IDF untuk melakukan generalisasi dengan lebih baik, menghasilkan peningkatan Macro *F1-Score* yang signifikan dari 0,7460 menjadi 0,7724 (±0,0264) pada Tugas-1A, dan dari 0,5471 menjadi 0,5798 (±0,0327) pada Tugas-1B. Hal ini dapat dijelaskan oleh sifat TF-IDF yang sangat bergantung pada kelengkapan kosakata (*vocabulary dependent*): meskipun tidak terlihat pada data validasi, semakin besar Korpus pelatihan yang digunakan, semakin kaya pemetaan kata secara statistik untuk menghadapi data baru yang tidak terduga[12].

4.6 Perbandingan dengan Peringkat Resmi (HASOC 2021)

Untuk menguji kinerja model yang diusulkan secara global, hasil terbaik dari penelitian ini dibandingkan dengan skor resmi para peserta pada papan peringkat HASOC 2021 kategori Bahasa Inggris. Perbandingan selengkapnya disajikan pada Tabel 7 dan Tabel 8.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 7. Perbandingan dengan peserta resmi tugas 1A HASOC 2021 bahasa inggris (tidak berpartisipasi secara resmi)

Peringkat	Nama Tim / Arsitektur	Metode	Macro Avg F1-Score Tugas-1A
1	NLP-CIC	*Large Transformer fine-tuning*	0.8305
2	HUNLP	GCN + TF-IDF	0.8215
3	neuro-utmn-thales	*Fine-tuned* Twitter-RoBERTa	0.8199
18	IMS-SINAI	BERT-*based Multi-Task Learning*	0.7947
*	RoBERTa+CatBoost (Ours)	*Frozen embedding.* + *Default* CatBoost	0.7924
19	SSN_NLP_MLRG	*Transformer ensemble*	0.7919
24	TAD	-	0.7776
*	TF-IDF+CatBoost(Baseline)	TF-IDF + *Default* CatBoost	0.7724
25	Beware Haters	LIME-*guided Ensemble* + TF-IDF	0.7722

Sebagaimana ditunjukkan pada Tabel 7, pendekatan RoBERTa + CatBoost *default* yang diusulkan secara teoritis berhasil mendekati peringkat ke-18 dan menggunguli peringkat ke-19 pada Tugas-1A dengan skor 0,7947, mengungguli SSN_NLP_MLRG dengan skor 0.7919. Pencapaian ini sangat signifikan mengingat tidak dilakukan penyempurnaan *(fine-tuning)* dan tidak diterapkan optimasi *hiperparameter* pada klasifikasi CatBoost [7]. Sebagian besar tim yang menduduki peringkat lebih tinggi mengandalkan sumber daya komputasi yang jauh lebih besar melalui penyempurnaan model *Transformer* berskala besar pada perangkat keras GPU khusus.

Satu-satunya tim yang menggunakan TF-IDF sebagai fitur utama di papan peringkat atas, HUNLP di peringkat ke-2, berhasil meraih skor 0,8215 dengan menggabungkan TF-IDF dan GCN *(Graph Convolutional Networks)*, yang membutuhkan kompleksitas model dan biaya komputasi yang jauh lebih besar [19]. Hal ini menunjukkan bahwa kombinasi TF-IDF sederhana + CatBoost *default* (0,7724) masih memiliki ruang untuk perbaikan yang signifikan, sementara RoBERTa *Embedding* saja sudah mampu bersaing hanya dengan menggunakan klasifikasi *default* yang siap pakai.

Tabel 8. Perbandingan dengan peserta resmi tugas 1B HASOC 2021 bahasa inggris (tidak berpartisipasi secara resmi)

Peringkat	Nama Tim / Arsitektur	Metode	Macro Avg F1-Score Tugas-1A
1	NLP-CIC	*Large Transformer fine-tuning*	0.6657
2	neuro-utmn-thales	HRF: BERTweet-*large*	0.6577
3	HASOC21rub	*Hybrid Representation Fusion*	0.6482
17	SATLab	-	0.6114
*	RoBERTa+CatBoost (ours)	*Frozen embed.* + *Default*	0.6113

 ISSN:2302-8149
e-ISSN:2540-9719

		CatBoost	
18	IRLab@IITBHU	*Transformer fine-tuning*	0.6093
22	Vishesh Gupta	*Ensemble* ML tradisional	0.5871
*	TF-IDF+CatBoost(Baseline)	TF-IDF + *Default* CatBoost	0.5798
23	MUM	PreCog IIIT Hyderabad	0.5771

Berdasarkan Tabel 8, pendekatan RoBERTa dengan CatBoost *default* yang diusulkan memperoleh Macro *F1-Score* sebesar 0.6113. Skor ini membuktikan bahwa sistem yang dikembangkan terdapat peningkatan pada tugas klasifikasi multi-kelas *(multiclass)*, menempatkannya hampir setara dengan peringkat ke-17 (SATLab dengan skor 0.6114) dan berhasil mengungguli tim-tim yang menggunakan ensambel Machine Learning tradisional

Menurut laporan resmi dari penyelenggara kompetisi HASOC 2021, klasifikasi berskala halus *(fine-grained classification)* pada Tugas-1B memang terbukti memberikan tantangan yang sangat tinggi[3]. Kesulitan ini berakar dari tingginya tingkat ambiguitas linguistik saat membedakan batas nuansa antara kelas ujaran kebencian *(Hate)*, kata kasar *(Profane)*, dan konten yang sekadar menyinggung *(Offensive)*. Hal ini turut dibuktikan oleh sangat rendahnya tingkat kesepakatan antar-anotator *(interrater agreement)* selama proses pelabelan data resmi untuk tugas tersebut[3].

5 Kesimpulan

Berdasarkan hasil eksperimen yang komprehensif dan analisis komparatif terhadap model-model klasifikasi ujaran kebencian pada dataset benchmark HASOC 2021 bahasa Inggris, dapat ditarik tiga kesimpulan utama. Pertama, keefektifan model hibrida RoBERTa dan CatBoost bawaan telah menunjukkan peningkatan. Penggunaan representasi semantik dari *embedding cardiffnlp/twitter-roberta-base-offensive* yang dibekukan dan diklasifikasikan menggunakan CatBoost bawaan tanpa optimasi parameter secara konsisten menghasilkan kinerja prediksi yang andal dan stabil. Hal ini menegaskan bahwa fitur keadaan tersembunyi *Transformer* sudah memiliki kualitas semantik yang cukup tinggi sehingga memungkinkan CatBoost beroperasi secara optimal dengan konfigurasi bawaan. Kedua, keunggulan RoBERTa dibandingkan TF-IDF telah terbukti secara empiris. *Embedding* RoBERTa secara konsisten melampaui batas kinerja maksimum TF-IDF klasik dalam kedua skenario klasifikasi (Tugas-1A ±0,0562, Tugas-1B ±0,0361). Kinerja RoBERTa terbukti jauh lebih unggul dan stabil dengan data pelatihan yang sama, dimana hal ini yang menegaskan keterbatasan mendasar TF-IDF dalam menangkap semantik kontekstual di luar statistik frekuensi kata. Ketiga, daya saing sistem ini terhadap tim-tim terdepan telah teruji. Sistem yang diusulkan, dengan skor 0,7924 (Tugas-1A) dan 0,6113 (Tugas-1B), mampu mengungguli peringkat ke-19 pada papan peringkat resmi HASOC 2021 Tugas-1A dan menggunguli peringkat ke-18 pada papan peringkat HASOC 2021 Tugas-1B, serta mampu bersaing dengan tim-tim internasional yang menggunakan sumber daya komputasi jauh lebih besar, semuanya tanpa penyempurnaan atau optimisasi *hiperparameter*.

Sebagai saran untuk pengembangan penelitian selanjutnya, terdapat beberapa aspek yang dapat dieksplorasi untuk menyempurnakan sistem deteksi ini. Pertama, mengingat tingkat kesulitan dan ambiguitas linguistik yang sangat tinggi pada klasifikasi berskala halus (Tugas-1B), penelitian mendatang disarankan untuk mengeksplorasi kombinasi fusi fitur (*hybrid representation fusion*) antara fitur Transformer dan leksikon tradisional untuk memetakan nuansa umpatan atau sarkasme dengan lebih presisi. Kedua, untuk mengatasi masalah ketidakseimbangan kelas yang lazim pada dataset media sosial, penerapan teknik penanganan data tidak seimbang (*imbalanced learning*) secara algoritmik, seperti penyesuaian bobot penalti kelas (*class weights*) pada CatBoost, sangat direkomendasikan. Terakhir, arsitektur model ini memiliki potensi besar untuk dikembangkan menjadi sistem deteksi yang memperhitungkan konteks percakapan eksternal (seperti riwayat

 ISSN:2302-8149
e-ISSN:2540-9719

interaksi pengguna atau rentetan balasan komentar) guna mengidentifikasi ujaran kebencian implisit yang tidak dapat dideteksi hanya dari teks kalimat tunggal

Referensi

[1] F. Alkomah and X. Ma, "*A Literature Review of Textual Hate Speech Detection Methods and Datasets*," Jun. 01, 2022, *MDPI*. DOI: 10.3390/info13060273.

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

[3] T. Mandl *et al.*, "*Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages under Creative Commons License Attribution 4.0 International (CC BY 4.0)*," 2021, [Online]. Available: http://ceur-ws.org

[4] S. Agustian, R. Saputra, and A. Fadhilah, "'*Feature Selection' with Pretrained-BERT for Hate Speech and Offensive Content Identification in English and Hindi Languages*," 2021. [Online]. Available: https://huggingface.co/surajp/RoBERTa-hindi-guj-san

[5] T. Wolf *et al.*, "*Transformers: State-of-the-Art Natural Language Processing*," 2020. [Online]. Available: https://github.com/huggingface/

[6] D. Quoc Nguyen, T. Vu, A. Tuan Nguyen, and V. Research, "*BERTweet: A Pre-Trained Language Model for English Tweets*," 2020. [Online]. Available: https://pypi.org/project/emoji

[7] E. Essa, K. Omar, and A. Alqahtani, "*Fake News Detection based on a Hybrid BERT and LightGBM Models*," *Complex and Intelligent Systems*, Vol. 9, No. 6, pp. 6581–6592, Dec. 2023, DOI: 10.1007/s40747-023-01098-0.

[8] Z. Guo, X. Wang, and L. Ge, "*Classification Prediction Model of Indoor PM2.5 Concentration using CatBoost Algorithm*," *Front. Built Environ.*, Vol. 9, 2023, DOI: 10.3389/fbuil.2023.1207193.

[9] R. A. Muhammad *et al.*, "Sistemasi: Jurnal Sistem Informasi Penerapan Metode *Naïve Bayes* dengan PSO pada Analisis Sentimen Pengguna Aplikasi X terhadap *Bitcoin Sentiment Analysis of X Application Users on Bitcoin using the Naïve Bayes Method Optimized with Particle Swarm Optimization (PSO)*," 2025. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[10] T. Mandl, S. Modha, M. Anand Kumar, and B. R. Chakravarthi, "*Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German*," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2020, pp. 29–32. DOI: 10.1145/3441501.3441517.

[11] J. M. Ahn, J. Kim, and K. Kim, "*Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-based CNN-LSTM for Harmful Algal Blooms Forecasting*," *Toxins (Basel).*, Vol. 15, No. 10, Oct. 2023, DOI: 10.3390/toxins15100608.

[12] A. Glazkova, M. Kadantsev, and M. Glazkov, "*Fine-Tuning of Pre-Trained Transformers for Hate, Offensive, and Profane Content Detection in English and Marathi*," 2021. [Online]. Available: https://pypi.org/project/tweet-preprocessor

[13] N. A. R. Putri and Ardiansyah, "Analisis Sentimen terhadap Kemajuan Kecerdasan Buatan di Indonesia menggunakan *BERT* dan *RoBERTa*," *Jurnal Sains dan Informatika*, Vol. 9, No. 2, pp. 136–145, Nov. 2023, DOI: 10.34128/jsi.v9i2.649.

[14] A. Kumar, P. K. Roy, and S. Saumya, "*An Ensemble Approach for Hate and Offensive Language Identification in English and Indo-Aryan Languages*," 2021. [Online]. Available: https://support.google.com/youtube/answer/2801939.

[15] G. B. Herwanto, A. M. Ningtyas, I. G. Mujiyatna, K. E. Nugraha, and I. N. Prayana Trisna, "*Hate Speech Detection in Indonesian Twitter using Contextual Embedding Approach*," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, Vol. 15, No. 2, p. 177, Apr. 2021, DOI: 10.22146/ijccs.64916.

[16] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, and M. Das, "*Exploring Transformer based Models to Identify Hate Speech and Offensive Content in English and Indo-Aryan Languages*,"

 ISSN:2302-8149
e-ISSN:2540-9719

2021. [Online]. Available: https://www.reuters.com/investigates/special-report/myanmar-facebook-hate

[17] Z. M. Farooqi, S. Ghosh, and R. R. Shah, "*Leveraging Transformers for Hate Speech Detection in Conversational Code-Mixed Tweets*," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.09986

[18] W. Yu and D. Kolossa, "*Hybrid Representation Fusion for Twitter Hate Speech Identification*," 2021. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

[19] N. Bölücü and P. Canbay, "*Hate Speech and Offensive Content Identification with Graph Convolutional Networks*," 2021. [Online]. Available: http://ceur-ws.org