BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

Klasifikasi Hate Speech dan Offensive Language Menggunakan BERT dan Support Vector Machine

Muhammad Tirta Syakban, Surya Agustian*, Muhammad Fikry, Muhammad Affandes

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Islam Negri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]11950115140@students.uin-suska.ac.id, [2,*]surya.agustian@uin-suska.ac.id, [3]muhammad.fikry@uin-suska.ac.id, [4]affandes@uin-suska.ac.id
Email Penulis Korespondensi: surya.agustian@uin-suska.ac.id

Abstrak−Ujaran kebencian dan bahasa ofensif merupakan permasalahan yang semakin kompleks pada media sosial dan membutuhkan pendekatan klasifikasi yang mampu menangkap konteks bahasa secara akurat. Meskipun model berbasis transformer yang di-*fine-tuning* secara end-to-end saat ini menjadi pendekatan dominan, eksplorasi penggunaan transformer sebagai *feature extractor* yang dikombinasikan dengan algoritma klasifikasi klasik masih relatif terbatas, khususnya dalam konteks benchmark seperti HASOC 2021.Penelitian ini bertujuan untuk menganalisis efektivitas pendekatan *feature-based transformer* dengan mengombinasikan embedding dari BERT dan RoBERTa dengan algoritma Support Vector Machine (SVM) dalam berbagai konfigurasi kernel, termasuk Linear, RBF, Polynomial, dan LinearSVC. Eksperimen dilakukan pada Sub-task A dan Sub-task B dengan membandingkan metode berbasis fitur tradisional (TF-IDF) dan embedding berbasis transformer. Hasil eksperimen menunjukkan bahwa embedding RoBERTa secara konsisten memberikan performa terbaik dibandingkan metode lainnya. Pada pengujian data test, kombinasi RoBERTa dan SVM menghasilkan performa yang kompetitif dibandingkan sistem lain pada HASOC 2021. Pada Sub-task B, model optimal mencapai nilai Macro F1-score sebesar 0.61, melampaui beberapa sistem berbasis BERT dan metode klasik. Temuan ini menunjukkan bahwa penggunaan embedding transformer sebagai representasi fitur yang dikombinasikan dengan SVM yang dioptimalkan dapat menjadi alternatif yang efektif terhadap pendekatan *fine-tuning*, khususnya dalam menghasilkan performa yang lebih stabil pada kondisi data yang tidak seimbang. Penelitian ini berkontribusi dalam menunjukkan potensi pendekatan *feature-based transformer* sebagai strategi klasifikasi yang fleksibel dan kompetitif dalam deteksi ujaran kebencian.

Kata Kunci: Klasifikasi; Ujaran Kebencian; Bahasa Ofensif; SVM; BERT

Abstract−Hate speech and offensive language have become increasingly complex problems on social media, requiring classification approaches that can effectively capture linguistic context. While transformer-based models with end-to-end fine-tuning have become the dominant approach, the use of transformers as fixed feature extractors combined with classical machine learning algorithms remains relatively underexplored, particularly in benchmark settings such as HASOC 2021. This study aims to investigate the effectiveness of a feature-based transformer approach by combining embeddings from BERT and RoBERTa with Support Vector Machine (SVM) classifiers using multiple kernel configurations, including Linear, RBF, Polynomial, and LinearSVC. Experiments were conducted on Sub-task A and Sub-task B by comparing traditional feature-based methods (TF-IDF) with transformer-based embeddings. The experimental results show that RoBERTa embeddings consistently outperform other feature extraction methods. On the test dataset, the combination of RoBERTa and SVM achieves competitive performance compared to other systems in HASOC 2021. In Sub-task B, the optimal model achieves a Macro F1-score of 0.61, outperforming several BERT-based and classical baseline systems.These findings demonstrate that using transformer embeddings as fixed feature representations combined with optimized SVM classifiers can serve as an effective alternative to fine-tuning approaches, particularly in achieving more stable performance under class imbalance conditions. This study contributes by highlighting the potential of feature-based transformer methods as a flexible and competitive strategy for hate speech and offensive language detection.

Keywords: Classification; Hate Speech; Offensive Language; SVM; BERT

1. PENDAHULUAN

Perkembangan pesat media sosial telah membuka ruang yang sangat luas bagi masyarakat untuk berinteraksi, menyampaikan pendapat, serta berbagi informasi secara cepat dan masif. Namun, fenomena ini juga diiringi dengan meningkatnya penyebaran ujaran kebencian (*hate speech*) dan bahasa ofensif (*offensive language*) yang berpotensi memicu konflik sosial, diskriminasi, serta kekerasan berbasis kebencian [1]. Dengan volume unggahan yang sangat besar setiap detik pada platform seperti X, proses moderasi konten secara manual menjadi tidak efisien dan sulit diterapkan secara konsisten. Oleh karena itu, pengembangan sistem otomatis untuk mendeteksi ujaran kebencian menjadi kebutuhan yang mendesak [2].

Dalam beberapa tahun terakhir, pendekatan berbasis *deep learning*, khususnya model transformer seperti BERT dan RoBERTa, telah menjadi metode dominan dalam klasifikasi teks. Model ini mampu membangun representasi kontekstual yang kaya melalui mekanisme *self-attention*, sehingga efektif dalam menangkap hubungan semantik yang kompleks dalam teks [3]. Keunggulan ini menjadikan transformer sebagai pendekatan utama dalam berbagai tugas *Natural Language Processing* (NLP), termasuk deteksi ujaran kebencian, sebagaimana juga tercermin dalam berbagai benchmark klasifikasi teks [4]. Kompetisi internasional seperti HASOC (*Hate Speech and Offensive Content Identification*) telah berkontribusi signifikan dalam menyediakan *benchmark* serta mendorong pengembangan berbagai pendekatan klasifikasi ujaran kebencian. Pada HASOC 2020, Mandl et al. [5] melaporkan bahwa pendekatan berbasis transformer seperti BERT, RoBERTa, dan XLM-R menunjukkan performa yang lebih unggul dibandingkan metode konvensional, khususnya pada bahasa Indo-Aryan dan data dengan karakteristik *code-mixing*. Hasil serupa juga dilaporkan oleh beberapa tim peserta seperti AI_ML_NIT_Patna [6] yang memanfaatkan *fine-tuning* model transformer

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

untuk meningkatkan performa klasifikasi. Pada HASOC 2021, model berbasis BERT tetap menunjukkan performa unggul dengan capaian *macro f1-score* yang tinggi pada berbagai subtask [7].

Pendekatan *multilingual* juga menunjukkan hasil yang menjanjikan. Model seperti XLM-RoBERTa mampu memanfaatkan transfer pengetahuan lintas bahasa sehingga meningkatkan performa klasifikasi pada bahasa dengan sumber daya terbatas [8], yang didukung oleh studi *cross-lingual transferability* [9]. Meskipun model transformer mendominasi performa terbaik, pendekatan berbasis fitur leksikal tetap menunjukkan relevansi yang kuat. Representasi seperti TF-IDF digunakan untuk menangkap distribusi kata dalam teks dan telah lama menjadi baseline yang stabil dalam klasifikasi teks [10]. Ketika dikombinasikan dengan algoritma Support Vector Machine (SVM), pendekatan ini mampu menghasilkan performa yang kompetitif pada berbagai tugas klasifikasi [11]. Secara teoretis, SVM memiliki keunggulan dalam membangun *decision boundary* optimal melalui prinsip *maximum margin*, sehingga efektif dalam menangani data berdimensi tinggi seperti teks [12]. Lebih lanjut, penelitian terbaru menunjukkan bahwa performa SVM dapat ditingkatkan melalui optimasi hyperparameter. Rofik et al. [13] menunjukkan bahwa penggunaan GridSearchCV mampu meningkatkan kinerja model SVM secara signifikan dalam tugas klasifikasi. Selain itu, pendekatan *hybrid* yang menggabungkan model bahasa seperti BERT sebagai *feature extractor* dengan SVM sebagai klasifikator juga mulai menunjukkan hasil yang konsisten. Anindya dan Kaesmetan [14] menunjukkan bahwa kombinasi BERT–SVM mampu meningkatkan akurasi pada analisis sentimen, sementara Iffa et al. [15] menunjukkan bahwa integrasi representasi BERT dengan SVM efektif meningkatkan performa pada kondisi *dataset* terbatas.

Pendekatan *feature-based* menjadi alternatif yang lebih ringan karena representasi kontekstual dari transformer dapat dimanfaatkan sebagai *feature extractor* tanpa melakukan pembaruan parameter model secara penuh memungkinkan proses klasifikasi dilakukan menggunakan algoritma *machine learning* klasik dengan kompleksitas pelatihan yang lebih rendah namun tetap memanfaatkan *semantic embedding* yang kaya dari transformer. Secara teoretis, *Support Vector Machine* (SVM) memiliki kemampuan generalisasi yang kuat pada data berdimensi tinggi melalui prinsip *maximum margin* dan *structural risk minimization*. Karakteristik ini menjadikan SVM sesuai untuk menangani representasi teks berbasis embedding maupun TF-IDF yang umumnya memiliki dimensi fitur besar dan *sparse* [16]. Sejumlah penelitian empiris juga menunjukkan bahwa SVM tetap mampu memberikan performa kompetitif pada berbagai tugas klasifikasi teks. Abdurohim et al. [17] melaporkan bahwa SVM dengan pembobotan TF-IDF mampu mencapai akurasi 96,82%, *precision* 94%, *recall* 92%, dan *F1-score* 93% pada klasifikasi komentar spam Instagram. Pada klasifikasi pencemaran nama baik di Twitter, Abdusyukur [18] menunjukkan bahwa model SVM memperoleh akurasi tertinggi sebesar 87,7% dengan kemampuan generalisasi yang stabil pada data latih dan data uji. Selain itu, Difandana dan Imaduddin [19] menunjukkan bahwa LinearSVC memiliki performa lebih unggul dibanding *Multinomial Naive Bayes* dengan selisih akurasi sebesar 8,7% pada klasifikasi ujaran kebencian dan teks abusif berbahasa Indonesia.

Kemudian berberapa pendekatan hybrid yang menggabungkan representasi fitur kuat dengan Support Vector Machine tetap mampu memberikan performa kompetitif pada berbagai tugas klasifikasi. Dalam domain medis, Wulandari et al [20]. membuktikan bahwa integrasi SVM dengan *Recursive Feature Elimination* dan ADASYN mampu mencapai akurasi 98.39%, menegaskan bahwa kualitas fitur dan optimasi preprocessing berperan signifikan terhadap performa model. Pada domain teks, penelitian JNATIA (2025) juga menunjukkan bahwa kombinasi SVM dengan teknik ekstraksi fitur dan penanganan *imbalance* seperti TF-IDF dan SMOTE mampu menghasilkan akurasi 90.82% pada klasifikasi sentimen media sosial, memperkuat bahwa SVM tetap relevan ketika didukung representasi fitur yang optimal [21]. Temuan-temuan tersebut menunjukkan bahwa pendekatan *hybrid* berbasis *transformer embedding* dan SVM masih memiliki relevansi yang kuat, khususnya pada skenario klasifikasi teks dengan *dataset* terbatas, distribusi data tidak seimbang.

Oleh karena itu, penelitian ini berfokus pada penggunaan algoritma *Support Vector Machine* (SVM) sebagai metode klasifikasi utama dalam mendeteksi ujaran kebencian dan bahasa ofensif, dengan memanfaatkan berbagai teknik ekstraksi fitur, termasuk BERT, RoBERTa, dan TF-IDF, serta penambahan data eksternal. Berbeda dengan penelitian sebelumnya, penelitian ini tidak menggunakan pendekatan BERT dan RoBERTa sebagai *fine-tuning*, melainkan mengeksplorasi penggunaan *feature extractor* yang dikombinasikan dengan berbagai konfigurasi SVM. Penelitian ini bertujuan untuk menganalisis sejauh mana pendekatan tersebut mampu menghasilkan performa yang kompetitif dibandingkan metode yang digunakan oleh peserta HASOC 2021, serta mengevaluasi pengaruh variasi *kernel* SVM terhadap kinerja klasifikasi.

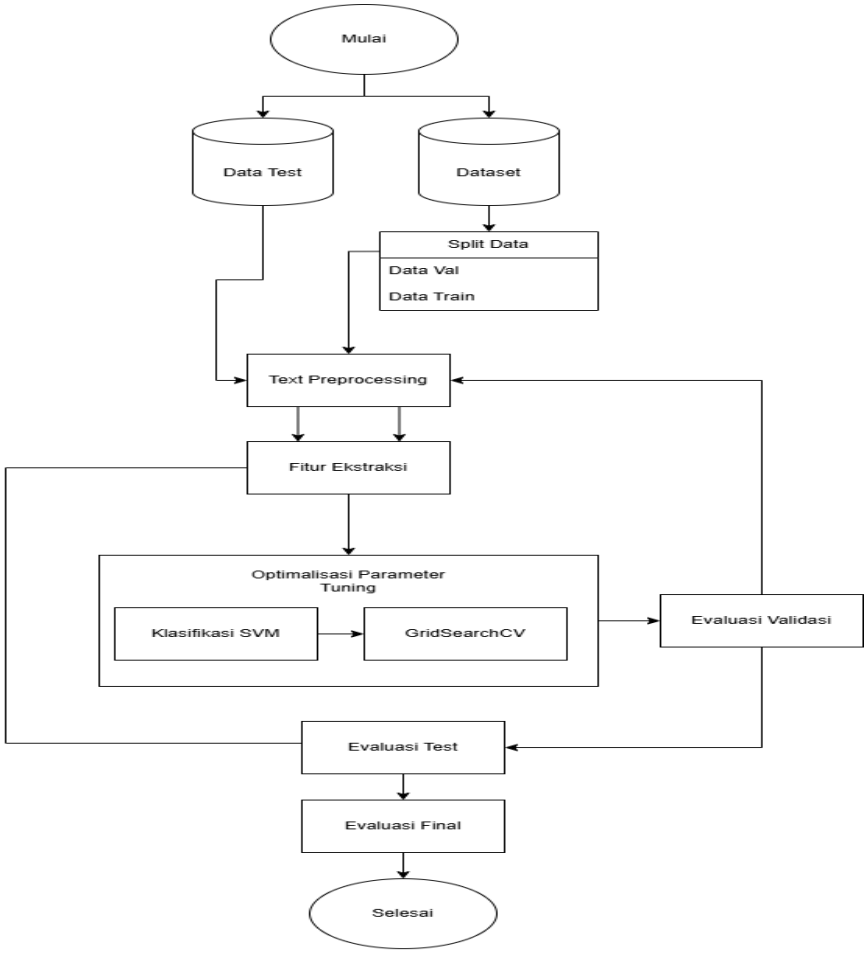
2. METODOLOGI PENELITIAN

2.1. Tahapan Penelitian

Penelitian dimulai dari dataset, yang dimana mencakup pengumpulan data dan seleksi data relevan dengan topik penelitian. Data kemudian diproses melalui tahap *processing* seperti penghapusan *mention* pengguna, URL, *normalization* teks, *Stemming*, Penyederhanakan huruf yang berulang dan *tokenizing*. Setelah tahapan *text processing*, dilakukan ekstraksi fitur dengan fitur ekstraksi yang berbeda seperti BERT, RoBERTa, dan TF-IDF dilanjutkan dengan klasifikasi dengan menggunakan algoritma *Support Vector Machine* (SVM) sebagai model utama. Setelah proses klasifikasi dilanjutkan dengan evaluasi model menggunakan metrik akurasi dan F1-score. Jika performa model belum optimal, maka penelitian dilakukan kembali dari tahap analisa data untuk memaksimalkan hasil performa model. Adapun

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

tahapan penelitian ini menggambarkan tahapan-tahapan yang harus dilalui untuk menyelesaikan penelitian, yang disajikan pada Gambar 1 berikut ini.



Gambar 1. Tahapan Penelitian

2.2. Dataset

Data yang digunakan dalam penelitian ini diperoleh dari FIRE Shared Task HASOC 2021, *Dataset* tersebut terdiri atas kumpulan teks yang diambil dari media sosial Twitter. FIRE Shared Task HASOC 2021 menyediakan 2 *task* klasifikasi. *Task* pertama adalah mengidentifikasi konten yang mengandung ujaran kebencian, bahasa ofensif, dan kata profan pada tweet berbahasa Inggris, Hindi, dan Marathi. Sementara itu, *task* kedua berfokus pada identifikasi ujaran kebencian dan bahasa ofensif dalam percakapan *tweet* yang menggunakan campuran bahasa (Inggris dan Hindi) [5]. Pada penelitian ini, kami hanya berfokus pada tugas pertama, yang dibagai menjadi 2 *sub-task* sebagai berikut.

a. *Sub-task* A

Sub-task A bertujuan untuk mengidentifikasi ujaran kebencian dan bahasa ofensif pada tweet berbahasa Inggris, Hindi, dan Marathi. Subtask ini merupakan klasifikasi biner tingkat kasar (*coarse-grained*), yang mengelompokkan *tweet* ke dalam dua kelas, yaitu:

1. HOF (*Hate, Offensive, and Profane)*: mencakup ujaran kebencian, bahasa ofensif, serta kata-kata profan.
2. NOT : mencakup konten yang tidak mengandung ujaran kebencian maupun bahasa ofensif.

b. *Sub-task* B

Subtask B merupakan tugas klasifikasi tingkat lanjut (*fine-grained*) untuk *tweet* berbahasa Inggris dan Hindi. Jika sebuah *tweet* telah diklasifikasikan sebagai HOF pada *Subtask* A, maka dilakukan klasifikasi lanjutan untuk menentukan kategori spesifiknya. *Tweet* tersebut kemudian diklasifikasikan menjadi salah satu dari tiga kelas berikut:

1. HATE (*Hate Speech*): *tweet* yang mengandung ujaran kebencian.
2. OFFN (*Offensive*): *tweet* yang mengandung bahasa ofensif
3. PRFN (*Profane*): *tweet* yang mengandung kata – kata profan
4. NONE: *tweet* yang tidak mengandung ujaran kebencian, bahasa ofensif dan kata profan

Dataset yang digunakan bersumber dari FIRE Shared Task HASOC 2021 task pertama yang berjumlah 3,843 teks tweet sebagai data latih dan 1281 teks tweet sebagai data uji. Dataset HASOC 2021 memuat tweet yang telah diklasifikasikan kedalam 2 kategori untuk sub-task A yaitu HOF dan NOT. Kemudian 4 kategori untuk sub-task B yaitu HATE, OFFN, PRFN, dan NONE pada Tabel 1.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

Tabel 1. label Dataset HASOC 2021

	Data latih						Data Uji
	Sub-task A		*Sub-task* B				
	HOF	NOT	HATE	OFFN	PRFN	NONE	
	2,501	1,342	683	622	1,196	1,342	1281
Total	3,843		3,843				1281

Dataset ini terdiri dari teks dalam format mentah yang berasal dari platform media sosial seperti Twitter. Bentuk teks pada *dataset* ini cenderung pendek, informal, dan mengandung berbagai elemen *noisy* seperti slang, *emoticon*, singkatan, hingga *grammar* yang tidak baku ditunjukkan pada Tabel 2. Hal ini membuat *dataset* HASOC 2021 sangat relevan untuk menguji model deteksi ujaran kebencian pada konteks dunia nyata.

Tabel 2. Data latih HASOC 2021

No	Teks *tweet*	Label 1	Label 2
1	@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 a	HOF	HATE
2	Technically that's still turning back the clock, dick head https://t.co/jbKaPJmpt1	HOF	OFFN
3	yeah when she's finally done w you you wanna pop back into her life fuck off	HOF	PRFN
4	@Warix_gay Can't expect God to do all the work	NOT	NONE

2.3. Pembagian Dataset

Pada penelitian ini, *dataset* dibagi menjadi dua, yaitu data *training* dan data *validation*, dengan proporsi 90% untuk *training* dan 10% untuk *validation.* Strategi pembagian ini diterapkan untuk memastikan bahwa model memiliki jumlah data yang memadai dalam proses pembelajaran, sekaligus menyediakan data yang representatif untuk mengevaluasi kinerja model selama proses pelatihan. Proses pembagian data dilakukan secara *stratified,* sehingga proporsi kelas pada data *training* dan data *validation* tetap mencerminkan distribusi kelas pada *dataset* asli. Data training pada penelitian ini terdiri dari dua sub-task, yaitu Sub-task A dan Sub-task B.

Pada Sub-task A, kategori HOF berjumlah 2.250 data dan NOT sebanyak 1.208 data dengan total 3.458 data. Sementara pada Sub-task B, kategori HATE berjumlah 615 data, OFFN sebanyak 560 data, PRFN sebanyak 1.076 data, dan NONE sebanyak 1.207 data dengan total 3.458 data. Data validation juga dibagi menjadi dua sub-task. Pada Sub-task A, kategori HOF berjumlah 251 data dan NOT sebanyak 134 data dengan total 385 data. Pada Sub-task B, kategori HATE berjumlah 68 data, OFFN sebanyak 62 data, PRFN sebanyak 120 data, dan NONE sebanyak 135 data dengan total 385 data. Pembagian ini menghasilkan proporsi yang tetap konsisten dengan distribusi awal *dataset*. *Data training* dimanfaatkan untuk proses pembelajaran model, termasuk ekstraksi fitur dan pelatihan algoritma klasifikasi. Sementara itu, *data validation* digunakan untuk memantau performa model selama proses pelatihan, mencegah terjadinya *overfitting*, serta sebagai acuan dalam pemilihan konfigurasi model terbaik sebelum dilakukan pengujian pada data uji.

2.4. Text Preprocessing

Proses text preprocessing dilakukan untuk menghilangkan noise dan menormalisasi teks sehingga lebih sesuai untuk proses ekstraksi fitur dan klasifikasi. Pada pendekatan berbasis fitur tradisional seperti TF-IDF, preprocessing dilakukan secara lebih intensif untuk mengurangi noise dan dimensi fitur, mengacu pada pendekatan yang digunakan oleh S. Ratan [12]. Sebaliknya, pada pendekatan berbasis transformer (BERT dan RoBERTa), preprocessing dilakukan secara minimal untuk mempertahankan konteks linguistik yang penting, mengingat model transformer telah dilatih pada data teks mentah yang mencakup berbagai bentuk noise seperti tanda baca dan simbol. Adapun tahapan-tahapan yang dilakukan adalah sebagai berikut:

a. Menghapus mention (*@username*)
   Semua kata yang diawali dengan simbol *@username* dihapus. Mention tidak memberikan makna semantik terhadap sentimen atau jenis ujaran, dan dapat menimbulkan bias terhadap identitas tertentu.
b. Menghapus URL atau Link
   URL sering muncul dalam tweet dan tidak merepresentasikan makna linguistik. URL dihapus menggunakan pola regular expression untuk menghilangkan pola http:// atau https://.
c. Menghapus token *retweet* (RT)
   Token *retweet* yang umum muncul pada data Twitter dihapus karena hanya berfungsi sebagai penanda retweet dan tidak mengandung informasi semantik yang relevan terhadap isi ujaran.
d. Perbaikan kesalahan emoji
   Pada pendekatan TF-IDF, emoji tidak dapat direpresentasikan secara efektif sebagai fitur berbasis frekuensi kata, sehingga dihapus untuk menjaga konsistensi representasi fitur.
e. Normalisasi huruf (*lowercasing*)
   Pada TF-IDF, Seluruh teks dikonversi menjadi huruf kecil (*lowercase*) untuk menghindari perbedaan representasi kata akibat variasi huruf kapital, seperti "*Hate*" dan "*hate*", yang seharusnya dianggap sebagai kata yang sama.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

f. Menghapus tanda baca
Seluruh tanda baca seperti titik, koma, tanda seru, dan simbol lainnya dihapus. Pada pendekatan TF-IDF, tanda baca tidak memiliki makna semantik dan dapat meningkatkan dimensi fitur secara tidak perlu.

g. *Stemming*
Pada TF-IDF, Setiap kata dalam teks dipangkas ke bentuk dasarnya (*stem*) menggunakan *Snowball Stemmer*. Proses ini bertujuan untuk mengurangi variasi bentuk kata, seperti *running, runs*, dan *ran* menjadi satu bentuk dasar, sehingga frekuensi kata dapat direpresentasikan secara lebih efektif.

2.5. Term Frequency–Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency merupakan salah satu metode ekstraksi fitur paling umum dalam pemrosesan bahasa alami (NLP). TF-IDF bekerja dengan mengukur pentingnya sebuah kata atau karakter dalam suatu dokumen relatif terhadap seluruh korpus.[22] Nilai Term Frequency (TF) menggambarkan frekuensi kemunculan kata dalam dokumen, sedangkan Inverse Document Frequency (IDF) menurunkan bobot kata yang terlalu sering muncul dalam banyak dokumen sehingga dianggap kurang informatif. Kombinasi keduanya menghasilkan representasi numerik yang efektif untuk tugas klasifikasi teks seperti *hate speech detection* [10].

Pada penelitian ini, TF-IDF diaplikasikan dalam dua bentuk n-gram: *word n-grams* dan *character n-grams*. *Word n-grams* menangkap informasi berbasis kata, yang berguna untuk memahami konteks semantik lokal. Sementara itu, *character n-grams* mampu mengenali pola morfologis, variasi ejaan, serta *noisy text* yang lazim muncul pada media sosial (misalnya slang, singkatan, atau kesalahan penulisan). Dalam implementasinya, *word n-grams* digunakan pada rentang 1–3 kata untuk menangkap *unigram* sampai *trigram*, sedangkan *character n-grams* menggunakan rentang 4–5 karakter yang efektif mendeteksi pola struktural dalam teks [11]. Kedua representasi ini kemudian diubah menjadi vektor numerik berdimensi tetap menggunakan *TfidfVectorizer* dari *scikit-learn* dan digunakan sebagai masukan bagi model klasifikasi berbasis SVM.

2.6. Bidirectional Encoder Representations from Transformers (BERT)

Pada penelitian ini, proses ekstraksi fitur dilakukan menggunakan arsitektur Bidirectional Encoder Representations from Transformers (BERT) sebagai ekstraksi fitur. BERT merupakan model pra-latih berbasis *Transformer* yang menghasilkan representasi kontekstual dua arah, sehingga setiap token dipengaruhi baik oleh konteks sebelum maupun sesudahnya [3]. Model yang digunakan adalah bert-base-multilingual-cased, yaitu varian BERT multibahasa yang dilatih pada korpus Wikipedia dari 104 bahasa dunia. Model ini bersifat *cased*, sehingga tetap mempertahankan perbedaan huruf besar dan kecil selama proses tokenisasi. Arsitektur model terdiri dari 12 layer *transformer*, 12 *attention heads*, dan *hidden size* sebesar 768 dengan total parameter sekitar 110 juta. Kualitas representasi model multibahasa ini didukung oleh kemampuan BERT dalam mentransfer pengetahuan lintas bahasa (*cross-lingual transfer*) sebagaimana dibahas dalam penelitian terdahulu [9]. Dalam proses ekstraksi fitur, setiap teks terlebih dahulu ditokenisasi menggunakan tokenizer *WordPiece* BERT, yang menambahkan token khusus [CLS] dan [SEP]. Model BERT dijalankan dalam mode inferensi dengan mengaktifkan keluaran seluruh *hidden states*. Penelitian ini memanfaatkan rata-rata dari empat layer terakhir BERT, karena lapisan-lapisan akhir terbukti menangkap representasi semantik yang lebih stabil. Selanjutnya, dua jenis representasi diekstraksi, yaitu vektor token [CLS] dan vektor hasil *mean pooling* berbasis *attention mask* pada seluruh token valid (tanpa *padding*). Kedua vektor tersebut kemudian digabungkan (*concatenation*) sehingga menghasilkan representasi fitur berdimensi 1536 untuk setiap dokumen.

2.7. Robustly Optimized BERT Approach (RoBERTa)

Selain BERT, ekstraksi fitur pada penelitian ini juga memanfaatkan RoBERTa sebagai model transformer berbasis encoder yang merupakan pengembangan dari BERT dengan optimasi pada strategi pre-training, meliputi penghapusan Next Sentence Prediction (NSP), penggunaan dynamic masking, pelatihan dengan batch yang lebih besar, serta peningkatan jumlah data latih sehingga menghasilkan representasi bahasa yang lebih robust [23]. Model yang digunakan adalah twitter-roberta-base-hate (cardiffnlp/twitter-roberta-base-hate), yaitu varian RoBERTa yang telah melalui domain-adaptive pretraining pada korpus Twitter dan dioptimalkan khusus untuk tugas hate speech serta offensive language detection dalam benchmark TweetEval, sehingga lebih sesuai untuk karakteristik teks media sosial yang penuh variasi linguistik, slang, emoji, dan noise dibanding model generic [24]. Proses ekstraksi fitur dilakukan tanpa *fine-tuning* dengan pendekatan *feature-based*. Setiap teks ditokenisasi menggunakan tokenizer RoBERTa berbasis byte-level BPE. Selanjutnya, representasi token diperoleh dari *hidden state* layer terakhir model karena layer ini merepresentasikan konteks semantik paling tinggi setelah seluruh proses self-attention, sehingga paling efektif untuk tugas klasifikasi[25]. Untuk membentuk representasi dokumen, penelitian ini mengekstraksi dua jenis vektor, yaitu vektor token pertama (<s>) yang berfungsi serupa dengan token [CLS] pada BERT, serta vektor hasil *mean pooling* berbasis *attention mask* pada seluruh token valid. Kedua vektor tersebut kemudian digabungkan menggunakan teknik *concatenation* sehingga menghasilkan representasi fitur berdimensi 1536 untuk setiap teks.

2.8. Klasifikasi Support Vector Machine

Support Vector Machine (SVM) merupakan suatu mesin algoritma yang digunakan untuk klasifikasi. Konsep dasar dari SVM yaitu mencari *hyperplane* dengan margin terbesar sehingga garis pemisah tersebut berada tepat dipembagi antara

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

kelas positif dan kelas negatif [14]. Pada penelitian ini, Support Vector Machine (SVM) dengan beberapa variasi *kernel* untuk mengevaluasi pengaruh kompleksitas model terhadap performa deteksi masing – masing kelas. Tiga *kernel* utama yang diuji adalah:

a. *Linear kernel*, yang cocok untuk fitur berdimensi tinggi seperti TF-IDF;
b. *Polynomial kernel*, yang mampu menangkap hubungan non-linear melalui pengaturan parameter degree, C, dan gamma;
c. *Radial Basis Function* (RBF) *kernel*, yang secara efektif memetakan data ke ruang fitur non-linear dengan parameter C dan gamma untuk mengatur margin dan sensitivitas model.

Selain itu, penelitian ini juga menguji *LinearSVC*, yaitu implementasi SVM linear berbasis *liblinear* yang lebih efisien untuk *dataset* besar. Beberapa nilai C dievaluasi untuk menentukan tingkat regularisasi terbaik. Eksperimen dengan berbagai *kernel* dan parameter dilakukan untuk mendapatkan konfigurasi yang optimal dalam memodelkan karakteristik bahasa pada *dataset* HASOC 2021

2.9. Parameter Tunning

Parameter tuning adalah proses penyesuaian dalam model untuk memperoleh performa optimal pada klasifikasi yang bertujuan mencari kombinasi parameter maksimal dalam metriks evaluasi pada validasi silang tanpa menyebabkan *overfitting* [26]. Pada penelitian ini metode *parameter tuning* dilakukan menggunakan metode *GridSearchCV*, metode ini digunakan untuk mencari kombinasi *hyperparameter* terbaik dengan skema pada tabel 3, seperti variasi *kernel*, nilai parameter C, gamma, serta degree untuk kernel polynomial [13]. Pencarian parameter dilakukan menggunakan *cross-validation* berserta *RepeatedStratifiedKFold* atau *StratifiedKFold* agar evaluasi stabil pada *dataset* yang tidak seimbang, sehingga proses optimasi dapat menangkap konfigurasi yang generalis. Selain itu, penelitian ini juga menggunakan LinearSVC sebagai pembanding, karena model ini dirancang untuk bekerja secara efisien pada fitur berdimensi tinggi seperti TF-IDF, serta memiliki kompleksitas yang lebih rendah dibandingkan SVC.

Tabel 3. Hyperparameter SVM

No.	Support Vector Machine	Hyperparameter		
		C	Gamma	Degree
1	Kernel Linear	1, 10, 100	-	-
2	Kernel RBF	1, 10, 100	1, 0.1, 0.01, 0.001	-
3	Kernel Polynomial	1, 10, 100	1, 0.1, 0.01, 0.001	2, 3, 4
4	LinearSVC	0.001, 0.01, 0.1, 1, 5, 10	-	-

2.10. Evaluasi

Evaluasi kinerja model dalam penelitian ini dilakukan menggunakan *confusion matrix*, yang menggambarkan distribusi prediksi model terhadap setiap kelas, meliputi *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Melalui *confusion matrix*, dapat dilihat pola kesalahan model, khususnya pada kasus ketidakseimbangan kelas. Dari nilai TP, FP, dan FN pada masing-masing kelas, dihitung metrik *Precision*, *Recall*, dan *f1-Score* secara per kelas sebagaimana dengan persamaan (1) [15]. Penelitian ini berfokus pada *macro f1-score*, yaitu nilai rata-rata F1 dari seluruh kelas tanpa mempertimbangkan proporsi jumlah data per kelas sebagaimana dengan persamaan (2) dengan k adalah jumlah kelas. Dengan demikian, setiap kelas memiliki bobot penilaian yang sama, sehingga *macro f1-score* lebih representatif untuk mengevaluasi performa model pada dataset yang tidak seimbang. Berikut adalah persamaan dari *F1-score* dan *macro F1-score*.

$$F1 - score = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (1)$$

$$Macro\ F1 - score = \frac{1}{K}\sum_{i=1}^{K} f1 - score_i \quad (2)$$

3. HASIL DAN PEMBAHASAN

3.1. Eksperimen Setup

Eksperimen Setup adalah rancangan dan konfiguraasi eksperimen yang digunakan untuk mengevaluasi performa model klasifikasi. Proses ini melibatkan berberapa tahap eksperimen dengan skenario sebagai berikut:

a. Eksplorasi fitur
   Tahap pertama dalam eksperimen adalah eksplorasi fitur, yang bertujuan untuk menganalisis pengaruh berbagai metode ekstraksi fitur terhadap performa model klasifikasi. Penelitian ini menggunakan tiga pendekatan utama dalam ekstraksi fitur, yaitu TF-IDF (*word n-gram* dan *character n-gram*), BERT (bert-base-multilingual-cased), dan RoBERTa (cardiffnlp/twitter-roberta-base-hate).
b. Penerapan Model Baseline
   Pada tahap ini, model dijalankan menggunakan konfigurasi standar tanpa penyesuaian *hyperparameter*. Ekstraksi fitur dilakukan menggunakan model *transformer*, kemudian diklasifikasikan menggunakan algoritma Support Vector

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

Machine (SVM) dengan parameter *default*, yaitu *kernel* RBF dengan nilai C = 1 dan gamma = *scale*. Evaluasi dilakukan pada *data validation* untuk memperoleh gambaran performa awal model pada masing-masing subtask. Hasil dari model baseline ini dijadikan pembanding untuk menilai peningkatan performa pada tahap optimasi.

c. Evaluasi Kombinasi Ekstraksi Fitur
Pada tahap ini, dilakukan pengujian terhadap berbagai kombinasi antara metode ekstraksi fitur dan konfigurasi model SVM. Setiap representasi fitur diuji menggunakan beberapa varian SVM, meliputi *kernel* linear, radial basis function (RBF), polynomial, serta LinearSVC. Eksperimen ini bertujuan untuk mengevaluasi sejauh mana kompleksitas kernel dan karakteristik fitur memengaruhi kemampuan model dalam memisahkan kelas pada tugas klasifikasi biner (*Sub-task* A) maupun multi-kelas (*Sub-task* B).

d. Parameter Tunning
Tahap terakhir dalam eksperimen adalah proses *parameter tuning* untuk memperoleh konfigurasi model yang optimal. Penyesuaian *hyperparameter* dilakukan menggunakan metode *GridSearchCV* dengan skema validasi silang berbasis *StratifiedKFold* atau *RepeatedStratifiedKFold*. Parameter yang dioptimasi meliputi nilai C, gamma, dan degree, tergantung pada jenis *kernel* SVM yang digunakan. Evaluasi performa dilakukan menggunakan metrik *macro F1-score* pada *data validation*. Model terbaik dipilih berdasarkan nilai *macro F1-score* tertinggi serta kestabilan performa antar *fold*.

3.2. Model Baseline

Model Baseline adalah model dasar untuk mendapatkan gambaran dasar klasifikasi serta aucan dasar peningkatan yang akan dilakukan dalam proses optimalisasi model. Model baseline meliputi, ekstraksi fitur menggunakan BERT dan klasifikasi model SVM *default* dimana kernel RBF dengan parameter C sama dengan 1 dan gamma *scale* untuk kedua *sub-task*. Hasil klasifikasi dari *Sub-task* A terhadap data validasi, yang diukur secara *macro-average* untuk F1-score, *accuracy*, *precision* dan *recall* secara berturut-turut adalah 0.68, 0.74, 0.73 dan 0.74. dan *Sub-task* B terhadap data validasi, yang diukur secara *macro-average* untuk F1-score, *accuracy*, *precision* dan *recall* secara berturut-turut adalah 0.55, 0.63, 0.61,dan 0.63. Kemudian model baseline ini diuji menggunakan data *test* untuk melihat kemampuan model dalam mengklasifikasikan data baru.

3.3. Proses Optimasi Model

Setelah melalui tahapan *text preprocessing*, penelitian ini melakukan proses optimalisasi model untuk memperoleh kinerja terbaik dalam klasifikasi. Optimalisasi dilakukan melalui dua pendekatan utama, yaitu pemilihan metode ekstraksi fitur dan pencarian konfigurasi *hyperparameter* terbaik untuk algoritma Support Vector Machine (SVM).

Tabel 4. Skenario Optimasi Ekstraksi Fitur dan SVM

ID	Ekstraksi Fitur	Support Vector Machine
A1	RoBERTa	Linear
A2	RoBERTa	RBF
A3	RoBERTa	Polynomial
A4	RoBERTa	LinearSVC
A5	BERT	Linear
A6	BERT	RBF
A7	BERT	Polynomial
A8	BERT	LinearSVC
A9	TF-IDF	Linear
A10	TF-IDF	RBF
A11	TF-IDF	Polynomial
A12	TF-IDF	LinearSVC

Berdasarkan pada Tabel 4, tahap ekstraksi fitur menggunakan tiga pendekatan berbeda. Pertama, digunakan representasi berbasis *transformer* dari model Roberta (cardiffnlp/twitter-roberta-base-hate). Kedua, digunakan pendekatan *feature extraction* berbasis TF-IDF dengan penambahan bentuk *word n-gram* maupun *character n-gram*. Ketiga, penelitian ini menguji embedding dari model BERT (bert-base-multilingual-cased) untuk melihat apakah representasi bahasa yang lebih umum dan berskala multilingual dapat memberikan performa lebih baik dibandingkan model yang spesifik terhadap domain media sosial. Setiap metode ekstraksi fitur tersebut kemudian diuji menggunakan algoritma SVM yang telah dioptimalkan melalui *GridSearchCV*. Seluruh kombinasi antara metode ekstraksi fitur dan model SVM diujikan secara terpisah untuk memperoleh model paling optimal. Kemudian model yang optimal pada masing kombinasi dievaluasi pada *data validation*, dengan metrik utama *macro F1-score* sesuai standar kompetisi dan akurasi sebagai pemisah.

Tabel 5. Hasil Optimasi *Sub-task* A terhadap Data Validasi

No.	ID	Ekstraksi fitur	SVM	Hyperparameter			Macro f1-score	Akurasi
				C	gamma	degree		
1	A2	RoBERTa	RBF	1	0.01	-	0.76	0.78

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

No.	ID	Ekstraksi fitur	SVM	Hyperparameter			Macro f1-score	Akurasi
				C	gamma	degree		
2	A3	RoBERTa	Polynomial	100	0.001	3	0.75	0.78
3	A4	RoBERTa	LinearSVC	0.001	-	-	0.75	0.77
4	A1	RoBERTa	Linear	1	-	-	0.75	0.76
5	A10	TF-IDF	RBF	10	1	-	0.72	0.77
6	A9	TF-IDF	Linear	1	-	-	0.72	0.75
7	A12	TF-IDF	LinearSVC	0.1	-	-	0.72	0.74
8	A6	BERT	RBF	1	0.01	-	0.70	0.72
9	A8	BERT	LinearSVC	0.001	-	-	0.70	0.71
10	A11	TF-IDF	Polynomial	100	0.1	3	0.69	0.72
11	A7	BERT	Polynomial	1	0.01	3	0.69	0.71
12	A5	BERT	Linear	1	-	-	0.68	0.70

Berdasarkan hasil eksperimen *sub-task* A yang ditampilkan pada tabel 5, ekstraksi fitur RoBERTa menghasilkan performa terbaik dibandingkan BERT dan TF-IDF. Kombinasi RoBERTa dengan SVM *kernel* RBF memberikan nilai *Macro F1-score* tertinggi sebesar 0.76 dengan akurasi 0.78, menunjukkan kemampuan model dalam menangani pola non-linear pada *embedding* berdimensi tinggi. Selain itu, LinearSVC dengan fitur RoBERTa juga menunjukkan performa yang kompetitif dengan *Macro F1-score* sebesar 0.75, meskipun masih berada di bawah *kernel* RBF. Hal ini mengindikasikan bahwa model linear tetap efektif, namun kurang fleksibel dibandingkan *kernel* non-linear. Sementara itu, penggunaan TF-IDF dan BERT menghasilkan performa yang lebih rendah dengan *Macro F1-score* masing-masing berada pada rentang 0.69 sampai dengan 0.72 dan 0.68 sampai dengan 0.70.

Tabel 6. Hasil Optimasi *Sub-task* B terhadap Data Validasi

No.	ID	Ekstraksi fitur	SVM	Hyperparameter			Macro F1-score	Akurasi
				C	gamma	degree		
1	A4	RoBERTa	LinearSVC	0.001	-	-	0.62	0.65
2	A2	RoBERTa	RBF	1	0.01	-	0.62	0.65
3	A12	TF-IDF	LinearSVC	0.1	-	-	0.60	0.65
4	A11	TF-IDF	Polynomial	10	1	2	0.60	0.65
5	A3	RoBERTa	Polynomial	1	0.01	3	0.60	0.64
6	A10	TF-IDF	RBF	100	1	-	0.59	0.65
7	A8	BERT	LinearSVC	0.01	-	-	0.57	0.63
8	A6	BERT	RBF	10	0.001	-	0.56	0.64
9	A9	TF-IDF	Linear	10	-	-	0.56	0.61
10	A1	RoBERTa	Linear	1	-	-	0.54	0.57
11	A7	BERT	Polynomial	10	1	3	0.52	0.58
12	A5	BERT	Linear	1	-	-	0.48	0.54

Berdasarkan hasil eksperimen *sub-task* B yang ditampilkan pada tabel 6, ekstraksi fitur RoBERTa kembali menunjukkan performa yang relatif lebih baik dibandingkan TF-IDF dan BERT. Nilai *Macro F1-score* tertinggi sebesar 0.62 diperoleh oleh dua skenario, yaitu RoBERTa dengan LinearSVC dan RoBERTa dengan SVM *kernel* RBF, dengan tingkat akurasi yang sama sebesar 0.65. Hasil ini menunjukkan bahwa baik model linear maupun non-linear masih mampu memberikan performa yang seimbang pada subtask ini, meskipun tingkat kompleksitas klasifikasi lebih tinggi dibandingkan *sub-task* A. Sementara itu, penggunaan TF-IDF menghasilkan performa menengah dengan *Macro F1-score* berada pada rentang 0.56 sampai dengan 0.60, sedangkan BERT menunjukkan performa terendah dengan nilai *Macro F1-score* berkisar antara 0.48 sampai dengan 0.57.

3.4. Hasil Evaluasi Data Test

Pengujian model ke *data test* dilakukan setelah mendapatkan model terbaik dari hasil evaluasi pada *data validation*. Pengujian ini bertujuan untuk menilai kemampuan model dalam mengklasifikasi pada data baru dan belum dilabel sehingga mendapatkan evaluasi performa model tersebut. Tabel 7 dibawah ini hasil pengujian model yang menampilkan metrik evaluasi pada masing – masing subtask untuk memberikan gambaran lengkap tentang performa tiap metode pada pengujian yang dilakukan dalam penelitian ini.

Tabel 7. Hasil Evaluasi *Data Test*

Sub -Task	Model	Ekstraksi fitur	SVM	Hyperparameter			Macro F1-score
				C	gamma	Degree	
A	Optimasi	RoBERTa	RBF	1	0.01	-	0.78
	Baseline	BERT	RBF	1	Scale	-	0.70
B	Optimasi 1	RoBERTa	LinearSVC	0.001	-	-	0.61
	Optimasi 2	RoBERTa	RBF	1	0.01	-	0.60

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

Sub -Task	Model	Ekstraksi fitur	SVM	Hyperparameter			Macro F1-score
				C	gamma	Degree	
	Baseline	BERT	RBF	1	Scale	-	0.54

Berdasarkan hasil pengujian pada *data test*, diperoleh peningkatan performa yang signifikan setelah dilakukan proses optimasi model pada masing-masing sub-task. Pada *Sub-task* A, model baseline menghasilkan nilai *Macro F1-score* sebesar 0.70. Setelah dilakukan optimasi menggunakan ekstraksi fitur RoBERTa dengan SVM *kernel* RBF (C = 1, gamma = 0.01), nilai *Macro F1-score* meningkat menjadi 0.78, yang menunjukkan bahwa proses optimasi berhasil meningkatkan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya. Pada *Sub-task B*, model baseline menghasilkan nilai *Macro F1-score* sebesar 0.54. Selanjutnya, optimasi pertama menggunakan RoBERTa dengan LinearSVC (C = 0.001) memberikan peningkatan performa yang cukup signifikan dengan nilai *Macro F1-score* sebesar 0.61. Sementara itu, optimasi kedua menggunakan RoBERTa dengan SVM *kernel* RBF (C = 1, gamma = 0.01) menghasilkan nilai *Macro F1-score* sebesar 0.60. Hasil ini menunjukkan bahwa pada *Sub-task* B, LinearSVC dengan fitur RoBERTa memberikan performa terbaik pada *data test*.

3.5. Analisa Kesalahan

Berdasarkan hasil evaluasi pada Sub-task A yang ditunjukkan pada Gambar 5, model menunjukkan performa yang relatif baik dalam mendeteksi kelas *hate/offensive (HOF)*, yang tercermin dari nilai *recall* yang lebih tinggi dibandingkan kelas *non-offensive (NOT)*. Hal ini mengindikasikan bahwa sebagian besar konten ofensif berhasil dikenali oleh model. Namun, nilai *recall* yang lebih rendah pada kelas NOT menunjukkan adanya kecenderungan model untuk mengklasifikasikan sebagian teks non-ofensif sebagai ofensif (*false positive*). Dengan kata lain, model cenderung lebih sensitif terhadap sinyal kebahasaan yang diasosiasikan dengan ujaran ofensif.

	precision	recall	f1-score	support
HATE	0.5394	0.6116	0.5732	224
NONE	0.7742	0.6460	0.7043	483
OFFN	0.4675	0.3692	0.4126	195
PRFN	0.6809	0.8443	0.7538	379
accuracy			0.6565	1281
macro avg	0.6155	0.6178	0.6110	1281
weighted avg	0.6588	0.6565	0.6516	1281

Gambar 2. Hasil Evaluasi Model Sub-task A

Pada Sub-task B, yang melibatkan klasifikasi multi-kelas, performa model menunjukkan ketidakseimbangan antar kelas. Kelas *profanity (PRFN)* memiliki nilai *recall* tertinggi, yang menunjukkan bahwa model mampu mengenali kata-kata kasar eksplisit dengan baik. Sebaliknya, kelas *offensive (OFFN)* memiliki nilai *recall* terendah, yang mengindikasikan bahwa banyak data pada kelas ini tidak berhasil diklasifikasikan dengan benar. Temuan ini menunjukkan adanya kesulitan model dalam membedakan antara kategori yang memiliki kemiripan semantik.

	precision	recall	f1-score	support
HOF	0.8340	0.8371	0.8355	798
NOT	0.7292	0.7246	0.7269	483
accuracy			0.7947	1281
macro avg	0.7816	0.7809	0.7812	1281
weighted avg	0.7944	0.7947	0.7946	1281

Gambar 3. Hasil Evaluasi Model Sub-task B

Secara umum, terdapat beberapa pola kesalahan utama. Pertama, model cenderung lebih mudah mengenali ujaran dengan ciri leksikal eksplisit (misalnya kata kasar) dibandingkan ujaran yang bersifat implisit atau kontekstual. Kedua, adanya kemiripan makna antara kelas seperti *OFFN* dan *HATE* menyebabkan model sulit membedakan batas antar kategori. Ketiga, model menunjukkan kecenderungan bias terhadap kelas tertentu, yang kemungkinan dipengaruhi oleh distribusi data yang tidak seimbang. Perlu dicatat bahwa analisis ini didasarkan pada metrik agregat (*precision* dan *recall*), sehingga tidak memberikan informasi detail mengenai arah kesalahan antar kelas. Namun demikian, pola-pola tersebut tetap memberikan indikasi yang cukup mengenai karakteristik kesalahan model.

3.6. Pembahasan

Setelah mendapatkan hasil optimasi dari masing – masing subtask, langkah berikutnya adalah membandingkan peforma model ini dengan metode lain yang tercantum di dalam *leaderboard* penelitian. Pada *leaderboard* memuat hasil dari metode yang diuji dalam klasifikasi oleh peneliti terdahulu yaitu *leaderboard* pertama dalam HASOC 2021, peneliti dari UIN Suska Riau yang berpartisipasi, dan peneliti yang menggunakan model SVM sebagai klasifikasinya. Tabel 8 dan tabel 9 dibawah ini adalah tabel pembandingan dari *Sub-task* A dan *Sub-task* B.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 1051-1061
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1061

Tabel 8. Perbandingan Sistem *Sub-task A*

Rank	Nama Tim	Sistem	Macro F1-score
1	NLP-CIC	RoBERTa-tw-large	0.8305
8	UINSUSKA	BERT+NN	0.8024
10	UMUTeam	BERT	0.8013
22	hate-busters	BERT-t10_run2	0.7894
34	S_Cube	TF-IDF + SVM	0.7563
	Baseline	BERT + SVM	0.7056
	Hasil Optimalisasi	RoBERTa + SVM	0.7812

Berdasarkan Tabel 8, model RoBERTa + SVM memperoleh Macro F1-score sebesar 0.7812, lebih tinggi dibandingkan pendekatan TF-IDF + SVM milik S_Cube sebesar 0.7563 dan model baseline sebesar 0.7056. Namun, performanya masih berada di bawah NLP-CIC yang menggunakan RoBERTa-tw-large dengan nilai 0.8305. Perbedaan ini menunjukkan bahwa penggunaan transformer berukuran besar yang *fine-tuning* secara end-to-end mampu menghasilkan representasi konteks yang lebih optimal dibandingkan pendekatan hybrid pada penelitian ini. Pada model yang diusulkan, RoBERTa hanya digunakan sebagai *feature extractor*, sedangkan proses klasifikasi dilakukan oleh SVM. Pendekatan tersebut membatasi proses penyesuaian representasi fitur terhadap objektif klasifikasi karena parameter *transformer* tidak sepenuhnya dioptimalkan bersama *classifier*. Akibatnya, model lebih sulit menangkap hubungan kontekstual yang kompleks pada ujaran kebencian dibandingkan arsitektur *transformer end-to-end.*

Tabel 9. Perbandingan Sistem *Sub-task B*

Rank	Nama Tim	Sistem	Macro F1-score
1	NLP-CIC	RoBERTa-tw-large	0.6657
5	UINSUSKA	BERT+NN	0.6417
10	UMUTeam	Ensemble model	0.6289
17	hate-busters	run1_bert	0.6096
25	S_Cube	TF-IDF + SVM	0.5739
	Baseline	BERT + SVM	0.5481
	Hasil Optimalisasi	RoBERTa + SVM	0.6110

Pada Tabel 9, model RoBERTa + SVM memperoleh Macro F1-score sebesar 0.6110 dan masih berada di bawah NLP-CIC dengan nilai 0.6657. Selisih performa pada Sub-task B lebih besar dibandingkan Sub-task A karena tugas multi-kelas memiliki kompleksitas klasifikasi yang lebih tinggi. Selain harus mendeteksi ujaran ofensif, model juga harus membedakan beberapa kategori dengan karakteristik linguistik yang mirip. Dalam kondisi tersebut, model *transformer* skala besar memiliki keunggulan karena mampu mempelajari dependensi semantik yang lebih kompleks melalui proses *fine-tuning end-to-end*. Sementara itu, penggunaan SVM sebagai classifier pada penelitian ini menyebabkan proses pembelajaran konteks menjadi terbatas pada *embedding* hasil ekstraksi RoBERTa. Meskipun demikian, hasil penelitian menunjukkan bahwa pendekatan *hybrid* RoBERTa + SVM tetap mampu memberikan performa kompetitif dengan arsitektur yang lebih sederhana dan biaya komputasi yang lebih rendah dibandingkan model *transformer* penuh.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan *feature-based* dengan memanfaatkan embedding RoBERTa yang dikombinasikan dengan Support Vector Machine (SVM) mampu menghasilkan performa yang kompetitif dalam klasifikasi ujaran kebencian dan bahasa ofensif. Hasil eksperimen menunjukkan bahwa pemilihan metode ekstraksi fitur memiliki pengaruh signifikan terhadap kinerja model, di mana embedding berbasis transformer memberikan representasi yang lebih efektif dibandingkan pendekatan berbasis fitur leksikal. Meskipun pendekatan *fine-tuning* transformer masih menghasilkan performa yang lebih tinggi, hasil penelitian ini menunjukkan bahwa kombinasi embedding statis dan SVM tetap menjadi alternatif yang layak, terutama dalam konteks eksplorasi metode klasifikasi yang lebih fleksibel. Namun demikian, penelitian ini juga mengidentifikasi bahwa ketidakseimbangan distribusi kelas (*class imbalance*) masih menjadi tantangan utama, khususnya pada klasifikasi multi-kelas. Hal ini terlihat dari ketidakmerataan performa antar kelas, yang menunjukkan bahwa model cenderung lebih optimal dalam mengenali kelas dengan distribusi data yang dominan. Selain itu, eksplorasi metode *fine-tuning ringan* pada model transformer juga dapat menjadi alternatif untuk meningkatkan sensitivitas model terhadap kelas minoritas.

REFERENCES

[1] N. Bölücü and P. Canbay, "Hate Speech and Offensive Content Identification with Graph Convolutional Networks," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021, pp. 44–51. [Online]. Available: http://ceur-ws.org

[2] R. Kumar, V. Gupta, and R. Pamula, "Hate Speech and Offensive Content Identification in English Tweets," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org

[3] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl.

[4] N. Azmi Verdikha, R. Habid, and A. Johar Latipah, "Analisis DistilBERT dengan Support Vector Machine (SVM) untuk Klasifikasi Ujaran Kebencian pada Sosial Media Twitter," *METIK JURNAL*, vol. 7, no. 2, pp. 101–110, Dec. 2023, doi: 10.47002/metik.v7i2.583.

[5] 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)," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org

[6] K. Kumari and J. P. Singh, "AI_ML_NIT_Patna @HASOC 2020: BERT Models for Hate Speech Identification in Indo-European Languages," in *Proceedings of the FIRE 2020 Working Notes*, CEUR-WS.org, 2020. [Online]. Available: http://ceur-ws.org

[7] S. Agustian, R. Saputra, and A. Fadhilah, "'Feature Selection' with Pretrained-BERT for Hate Speech and Offensive Content Identification in English and Hindi Languages," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021. [Online]. Available: https://huggingface.co/surajp/RoBERTa-hindi-guj-san

[8] M. Bhatia *et al.*, "One to Rule Them All: Towards Joint Indic Language Hate Speech Detection," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org

[9] M. Artetxe, S. Ruder, and D. Yogatama, "On the Cross-lingual Transferability of Monolingual Representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 4623–4637. [Online]. Available: https://github.

[10] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset," in *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[11] Y. Hacohen-Kerner and M. Uzan, "Detecting Offensive Language in English, Hindi, and Marathi using Classical Supervised Machine Learning Methods and Word/Char N-grams," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021. [Online]. Available: http://www.icra.org/

[12] S. Ratan, S. Sinha, and S. Singh, "SVM for Hate Speech and Offensive Content Detection," in *Proceedings of the FIRE 2021 Working Notes*, CEUR-WS.org, 2021.

[13] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetiyo, and M. A. Muslim, "Optimization of SVM and Gradient Boosting Models Using GridSearchCV in Detecting Fake Job Postings," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419–430, Mar. 2024, doi: 10.30812/matrik.v23i2.3566.

[14] F. S. Anindya and Y. R. Kaesmetan, "Implementasi Metode BERT DAN SVM pada Analisis Sentimen Game Genshin Impact," *Jurnal Manajamen Informatika Jayakarta*, vol. 5, no. 1, p. 52, Feb. 2025, doi: 10.52362/jmijayakarta.v5i1.1781.

[15] M. R. Iffa, S. Agustian, N. Safaat, and M. Irsyad, "Peningkatan Kinerja Support Vector Machine Menggunakan Model Bahasa BERT untuk Klasifikasi Sentimen dengan Dataset Terbatas," *ZONAsi : Jurnal Sistem Informasi*, vol. 7, pp. 422–432, 2025, doi: https://doi.org/10.31849/zn.v7i2.26847.

[16] K. Hadi and E. Utami, "Analisis K-NN Dengan Integrasi BOW, TF-IDF, Dan N-Grams untuk Klasifikasi Ujaran Kebencian pada Twitter," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 4, pp. 2971–2983, Nov. 2025, doi: 10.29100/jipi.v10i4.6694.

[17] U. Abdurohim *et al.*, "Implementasi Algoritma Support Vector Machine (SVM) untuk Klasifikasi Komentar Spam pada Instagram," *Jurnal Teknologi Informasi dan Komunikasi*, vol. 13, no. 1, pp. 13–19, 2024, doi: 10.58761/jurtikstmikbandung.v13i1.1319.

[18] F. Abdusyukur, "Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Pencemaran Nama Baik di Media Sosial Twitter," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 1, pp. 73–82, 2023, doi: https://doi.org/10.34010/komputa.v12i1.9418.

[19] R. Difandana and I. Imaduddin, "Analisis Komparatif Algoritma Naive Bayes dan Support Vector Machine dalam Klasifikasi Ujaran Kebencian dan Teks Abusif Berbahasa Indonesia," *FON Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. 22, pp. 267–278, 2026, doi: 10.25134/fon.v22i1.471.

[20] C. Wulandari, L. Afrianty, E. Budianita, and S. K. Gusti, "Thyroid Disease Classification Using Support Vector Machine and Recursive Feature Elimination Method," *bit-Tech*, vol. 8, no. 2, pp. 2948–2960, Dec. 2025, doi: 10.32877/bt.v8i2.3454.

[21] A. A. P. Putra and G. A. G. A. K. Kadyanan, "Klasifikasi Citra Jamur Menggunakan SVM dengan PCA Berbasis Ekstraksi Fitur Hibrida," *Jurnal Nasional Teknologi Informasi dan Aplikasinya*, vol. 4, no. 2, pp. 243–252, 2026, doi: 10.24843/JNATIA.2026.v04.i02.p02.

[22] J. Rama Dani, "Analisis Sentimen Komentar YouTube terhadap Kenaikan Tunjangan DPR RI menggunakan Naïve Bayes, SVM, dan Random Forest," *Technology and Science (BITS)*, vol. 7, no. 3, pp. 1512–1524, 2025, doi: 10.47065/bits.v7i3.8513.

[23] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, Jul. 2019, doi: 10.48550/arXiv.1907.11692.

[24] L. P. Vecchi, A. De Souza Britto, E. Cabrera Paraiso, and R. Menelau Cruz, "HARM: Learning Hate-Aware Reward Model for Evaluating Natural Language Explanations of Offensive Content," in *Findings of the Association for Computational Linguistics: EACL 2026*, Association for Computational Linguistics (ACL), 2026, pp. 4393–4431. doi: https://doi.org/10.18653/v1/2026.findings-eacl.230.

[25] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Oct. 2020, pp. 1644–1650. doi: 10.18653/v1/2020.findings-emnlp.148.

[26] D. Ismunandar, M. R. Firdaus, and Y. Alkhalifi, "Penerapan Hyperparameter Machine Learning dalam Prediksi Gagal Pinjam," *INTI Nusa Mandiri*, vol. 19, no. 1, pp. 62–70, Aug. 2024, doi: 10.33480/inti.v19i1.5612.