Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

Analisis Sentimen Terhadap Cyberbullying di Twitter (X) Menggunakan Improved Word Vectors dan Bert

Madya Dharma Nusantara*, Fajri Rakhmat Umbara, Puspita Nurul Sabrina

Fakultas Sains Dan Informatika, Program Studi Teknik Informatika, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email: [1,*]madyadharma21@if.unjani.ac.id, [2]fajri.rakhmat@lecture.unjani.ac.id, [3]puspita.sabrina@lecture.unjani.ac.id
Email Penulis Korespondensi: madyadharma21@if.unjani.ac.id
Submitted: **11/07/2025**; Accepted: **01/09/2025**; Published**: 02/09/2025**

Abstrak−Text mining merupakan salah satu pendekatan penting dalam menganalisis data teks, khususnya untuk mendeteksi sentimen negatif seperti cyberbullying di media sosial. Twitter (X), sebagai platform terbuka, sering kali menjadi ruang berkembangnya ujaran kebencian dan perilaku melecehkan yang terekam dalam bentuk teks. Penelitian ini bertujuan untuk meningkatkan performa model klasifikasi sentimen pada data Twitter (X) melalui penggabungan metode Improved Word Vector (IWV) dan Bidirectional Encoder Representations from Transformers (BERT), yang dievaluasi menggunakan metrik precision, recall, dan F1-score. Dataset yang digunakan terdiri dari 9.874 tweet berbahasa Indonesia yang telah dilabeli ke dalam tiga kategori: Hate Speech (HS), Abusive, dan Netral. Data ini bersumber dari penelitian terdahulu dan merupakan hasil anotasi ulang dari dataset asli berjumlah 13.169 tweet. IWV dibentuk dari kombinasi fitur Word2Vec, GloVe, POS tagging, dan leksikon emosi yang dirancang untuk memperkaya representasi kata secara makna. Proses preprocessing dilakukan melalui beberapa tahapan penting, yaitu tokenisasi, filtering, stemming/lemmatization, dan normalisasi. Hasil ekstraksi IWV kemudian digabungkan dengan embedding BERT melalui teknik concatenation untuk menghasilkan representasi vektor berdimensi tinggi. Evaluasi dilakukan menggunakan metrik precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model gabungan IWV+BERT mampu menghasilkan performa yang lebih baik dibandingkan penggunaan BERT secara tunggal. Penggunaan data yang telah diseimbangkan melalui teknik balancing turut berkontribusi dalam peningkatan akurasi, dengan capaian nilai akurasi tertinggi sebesar 91%. Temuan ini mengindikasikan bahwa integrasi fitur representasi kata dari IWV dan konteks kalimat dari BERT mampu meningkatkan efektivitas text mining dalam analisis sentimen terkait cyberbullying di media sosial.

Kata Kunci: Text Mining; Analisis Sentimen; Cyberbullying; IWV; BERT

Abstract−Text mining is an important approach in analyzing text data, particularly for detecting negative sentiments such as cyberbullying on social media. Twitter (X), as an open platform, often serves as a space for the proliferation of hate speech and abusive behavior recorded in text form. This study aims to improve the performance of sentiment classification models on Twitter (X) data by combining the Improved Word Vector (IWV) and Bidirectional Encoder Representations from Transformers (BERT) methods, evaluated using precision, recall, and F1-score metrics. The dataset used consists of 9,874 Indonesian-language tweets labeled into three categories: Hate Speech (HS), Abusive, and Neutral. This data is sourced from previous research and is the result of re-annotation of the original dataset of 13,169 tweets. IWV is formed from a combination of Word2Vec, GloVe, POS tagging, and emotion lexicon features designed to enrich word representation semantically. The preprocessing process is carried out through several important stages, namely tokenization, filtering, stemming/lemmatization, and normalization. The IWV extraction results were then combined with BERT embedding through concatenation to produce high-dimensional vector representations. Evaluation was performed using precision, recall, and F1-score metrics. The test results showed that the combined IWV+BERT model was able to produce better performance than BERT alone. The use of data that has been balanced through balancing techniques also contributed to the improvement in accuracy, with the highest accuracy value reaching 91%. This finding indicates that the integration of word representation features from IWV and sentence context from BERT can improve the effectiveness of text mining in sentiment analysis related to cyberbullying on social media

Keywords: Text Mining; Sentiment Analysis; Cyberbullying; IWV; BERT

1. PENDAHULUAN

Media sosial telah menjadi ruang digital yang memungkinkan penggunanya untuk berbagi ide, pendapat, dan perasaan secara terbuka melalui berbagai platform, seperti video, foto, blog, permainan daring, jejaring bisnis, hingga kolom komentar sosial [1]. Namun, di balik kemudahan dan kebebasan berkomunikasi yang ditawarkan, muncul pula berbagai risiko negatif, salah satunya adalah cyberbullying. Tindakan ini mencakup perilaku melecehkan, mengancam, atau mempermalukan individu secara daring melalui media sosial, pesan teks, maupun forum diskusi [2]. Dalam satu dekade terakhir, Twitter dan berbagai platform media sosial lainnya mengalami pertumbuhan yang sangat pesat. Fitur seperti anonimitas pengguna dan kebebasan dalam menyampaikan pendapat menjadikan media sosial sebagai tempat yang subur bagi berkembangnya ujaran kebencian. Twitter, misalnya, tercatat memiliki sekitar 300 juta pengguna aktif setiap bulannya, menjadikannya salah satu platform terpopuler saat ini. Meskipun memiliki peran penting dalam komunikasi digital, Twitter juga kerap menjadi medium penyebaran ujaran kebencian [3]. Oleh karena itu, penting untuk melakukan analisis terhadap konten media sosial guna mengidentifikasi potensi ujaran negatif yang dapat berdampak pada kesehatan mental dan sosial pengguna.

Komentar di platform Twitter (X) umumnya berbentuk teks, sehingga diperlukan penerapan analisis *Text mining* untuk memahami isi dan makna yang terkandung di dalamnya. *Text mining* bertujuan untuk mengekstraksi informasi dari kumpulan dokumen teks serta mendukung proses penemuan pengetahuan dari koleksi dokumen yang besar dan tidak terstruktur. Secara umum, sumber data untuk *text mining* biasanya berasal dari teks dengan format yang tidak teratur atau paling tidak semi-terstruktur, seperti komentar pengguna di media sosial, ulasan produk, atau

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

berita daring [4]. *Text mining* merupakan teknik yang mampu mengolah data teks guna menghasilkan informasi yang berkualitas tinggi. Selain itu, metode ini dapat digunakan untuk menganalisis sentimen dalam suatu kalimat secara cepat, sehingga mendukung proses pengambilan informasi yang relevan dan bernilai [5]. *Text mining*, atau yang juga dikenal sebagai analisis teks, merupakan teknik yang memanfaatkan pemrosesan bahasa alami (NLP) untuk mengekstraksi informasi dari data teks tidak terstruktur. Teknik ini bertujuan untuk mengidentifikasi opini, emosi, evaluasi, sikap, hingga penilaian yang terkandung dalam teks, baik yang berkaitan dengan layanan, individu, organisasi, maupun kegiatan tertentu. Dalam konteks media sosial, text mining berperan penting dalam mengenali pola sentimen dan potensi interaksi negatif seperti ujaran kebencian atau cyberbullying, sehingga dapat digunakan untuk membangun sistem deteksi otomatis dan menciptakan lingkungan digital yang lebih sehat [6].

Pada penelitian [7], [8], [9] dijelaskan bahwa tahapan preprocessing memegang peranan yang sangat penting dalam proses klasifikasi teks, karena secara langsung dapat memengaruhi nilai akurasi dari model yang digunakan. Pemilihan dan penerapan tahapan preprocessing yang tepat, seperti tokenisasi, normalisasi, serta penghapusan kata-kata yang tidak relevan, akan berdampak positif terhadap kualitas fitur yang dihasilkan dan, pada akhirnya, meningkatkan akurasi prediksi model. Oleh karena itu, preprocessing tidak hanya menjadi langkah awal dalam proses data mining, tetapi juga merupakan salah satu faktor kunci dalam menentukan keberhasilan analisis data secara keseluruhan.

Tingkat akurasi dalam analisis sentimen menggunakan metode Bidirectional Encoder Representations from Transformers (BERT) pada berbagai penelitian sebelumnya umumnya menghasilkan nilai rata-rata di bawah 74,69% [1] , [10], [11], [12], [13]. Misalnya, penelitian berjudul "Analisis Sentimen terhadap Perundungan Siber pada Twitter menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT)" (2023) menunjukan akurasi sebesar 81% [1]. Penelitian lainnya, "Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024" (2024), menghasilkan akurasi 76,84% [10]. Sementara itu, "Analisis Sentimen KUHP Baru Pada Data Twitter Menggunakan Model BERT" (2022), mencatat akurasi sebesar 81% [11]. Dan selanjutnya "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers" (2020) menunjukan akurasi 73% [12]. Adapun penelitian "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (PSE) Menggunakan Algoritma Bidirectional Encoder Representasi From Transformers (BERT)" (2022), menghasilkan akurasi 69% [13].

Meskipun angka tersebut dapat dikategorikan cukup baik dan menunjukkan bahwa BERT memiliki performa yang menjanjikan dalam tugas-tugas klasifikasi teks, hasil tersebut masih menyisakan ruang untuk perbaikan, khususnya ketika model diterapkan pada permasalahan text mining yang lebih kompleks, seperti dalam deteksi ujaran kebencian atau cyberbullying. Terbatasnya pemahaman model terhadap makna kata yang bersifat semantik eksplisit atau hubungan gramatikal menjadi salah satu faktor yang menghambat pencapaian akurasi lebih tinggi. Selain itu, variasi akurasi yang ditemukan dalam berbagai penelitian juga menunjukkan bahwa performa BERT sangat dipengaruhi oleh karakteristik data serta tahapan preprocessing yang digunakan.

Sebagai solusi, penggabungan representasi kata dari Improved Word Vector (IWV) dan BERT dapat meningkatkan pemahaman model terhadap teks secara lebih mendalam dan kaya. IWV membantu menangkap makna kata secara semantik dan struktural, sementara BERT unggul dalam memahami konteks kalimat secara dinamis. Kombinasi keduanya memungkinkan model untuk memperoleh informasi dari dua sisi makna kata dan konteks penggunaannya sehingga analisis sentimen dapat dilakukan dengan lebih akurat dibandingkan jika hanya menggunakan salah satu metode secara tunggal. Metode IWV digunakan untuk meningkatkan akurasi representasi kata yang sebelumnya telah dilatih. Melalui pendekatan ini, embedding kata disempurnakan dengan menambahkan informasi semantik dan gramatikal, sehingga lebih mampu menangkap makna yang relevan dan kontekstual. Penelitian sebelumnya [14] menunjukkan bahwa IWV efektif dalam meningkatkan kualitas representasi kata, khususnya dalam berbagai tugas pemrosesan bahasa alami seperti klasifikasi dan analisis sentimen.

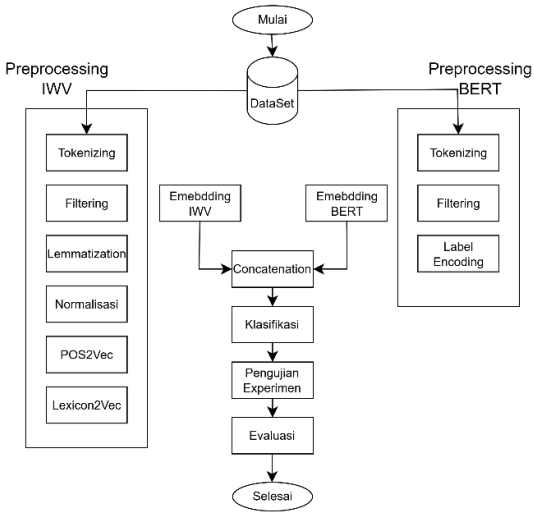
Penelitian ini bertujuan untuk mengevaluasi kontribusi penggabungan Improved Word Vector dengan model BERT dalam meningkatkan akurasi analisis sentimen. Eksperimen dilakukan untuk melihat sejauh mana penggabungan IWV dapat memperbaiki performa klasifikasi dibandingkan penggunaan BERT secara tunggal. Dataset yang digunakan dalam penelitian ini mengacu pada dataset dari studi sebelumnya [15], dan hasil yang diperoleh akan dibandingkan dengan penelitian terdahulu, khususnya pada [1]. Beberapa studi yang relevan terkait analisis sentimen dalam konteks cyberbullying di Indonesia juga telah dilakukan sebelumnya, antara lain oleh [1], [2] , [16], dan [17], yang menjadi referensi penting dalam pengembangan dan pembandingan hasil pada penelitian ini.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Proses penelitian ini dilaksanakan melalui serangkaian tahapan yang dirancang secara sistematis dan terstruktur guna memastikan bahwa setiap langkah yang ditempuh sesuai dengan tujuan yang telah ditetapkan. Rangkaian langkah-langkah tersebut dapat dilihat secara jelas pada Gambar 1, yang menyajikan alur kerja penelitian dari awal hingga akhir. Setiap tahapan memiliki peran yang signifikan dalam mendukung keberhasilan proses analisis, mulai dari pengumpulan dan praproses data hingga tahap klasifikasi dan evaluasi model.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968



Gambar 1. Alur Penelitian

2.2 Data Set

Data yang digunakan dalam penelitian ini merupakan data non-kormesil yang diperoleh dari platform GitHub melalui tautan tercantum dalam referensi [15]. Dataset ini awalnya terdiri dari 13.169 data berupa cuitan dari media sosial Twitter (X) dalam bahasa Indonesia, yang masing-masing telah diberi label oleh peneliti sebelumnya. Dataset asli memiliki dua belas kategori label, namun dalam penelitian ini hanya digunakan dua kategori, yaitu HS (Hate Speech) dan Abusive. Untuk keperluan klasifikasi yang lebih seimbang dan terfokus, peneliti juga menambahkan satu label tambahan, yaitu Netral, yang diperoleh melalui proses penyaringan ulang dan anotasi manual oleh peneliti, dengan mengidentifikasi data yang tidak termasuk dalam dua kategori utama yang disebutkan sebelumnya. Setelah proses penyaringan, pemilihan label, dan penyeimbangan data, diperoleh jumlah akhir sebanyak 9874 data. Penetapan label dan validasi data mengacu pada hasil diskusi dan kesepakatan dalam Focus Group Discussion (FGD) yang melibatkan peneliti dan staf dari Direktorat Tindak Pidana Siber Bareskrim Polri, guna menjamin konsistensi dan kesesuaian data dengan konteks hukum dan sosial di Indonesia.

2.3 Pra-pemrosesan Data

a. Tokenazing
   Tahap *tokenizing* atau tokenisasi dimana pemotongan *input* berdasarkan dari kalimat menjadi kata-kata terpisah.
b. Filtering
   Tahap *filtering* atau penyaringan dimana pengambilan kata-kata penting diambil hasil tokenisasi. Penyaringan dapat menggunakan algoritma *stop list* (menghapus kata-kata kurang penting) atau *word list* (menyimpan kata-kata penting).
c. Lematization
   Tahap l*emmatization* mengubah kata menjadi kata dasar dengan menghapus konjungsinya berdasarkan kamus.
d. Normalisasi
   Tahap normalisasi teks memperbaiki atau menyamakan kata-kata dalam teks sehingga sesuai dengan format standar.
e. Part-of-Speech (POS) Tagging
   Tahapan POS memberikan konteks tambahan pada kata-kata berdasarkan peran gramatikalnya.
f. Lexicon
   Tahap ini dibangun basis data yang memuat kata atau frasa yang telah dilabeli dengan atribut tertentu untuk merepresentasikan karakteristik linguistik atau domain spesifik.
g. Label Encoding
   Tahap ini merupakan salah satu teknik praproses data yang digunakan untuk mengubah nilai kategori (biasanya berupa data teks atau string) menjadi representasi numerik (angka).

2.4 IWV

Improved Word Vectors (IWV) adalah metode untuk meningkatkan akurasi embedding kata yang sudah dilatih sebelumnya (pre-trained) dalam analisis sentiment. Metode IWV ini menggabungkan teknik Part-of-Speech (POS) tagging, pendekatan berbasis leksikon, dan metode Word2Vec/GloVe untuk menghasilkan representasi vektor kata yang lebih akurat dan relevan dengan konteks sentiment[14].

a. Word2Vec
   Dalam penelitian ini saya menggunakan 300 corpus Word2Vec yang berasal dari *genism.downloader* dari google pada korpus Google News. Penelitian ini, memanfaatkan model Skip-gram dari Word2Vec, yang dimana dikenal

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

sebagai salah satu arsitektur utama dalam representasi kata. Skip-gram memiliki kemiripan konseptual dengan model CBOW (Continuos Bag-of-Words), namun dengan pendekataan yang berlawanan. Dimana jika pada CBOW memprediksi kata target dari konteks sekitarnya, Skip-gram memprediksi kata-kata konteks berdasarkan kata target saat ini. Dengan kata lain, model ini memaksimalkan peluang klasifikasi sebuah kata kontes berdasarkan kata targetnya. Tujuannya adalah untuk mencari representasi vektor kata yang memungkikan prediksi kata-kata di sekitarnya dengan baik. Dimana pada penelitian [18] menjelaskan bahwa model Skip-gram lebih baik dari model CBOW dengan menggunakan data set yang lebih kecil dan untuk kata-kata yang jarang. Berikut adalah rumus yang digunakan:
Fungsi Objektif Skip-Gram

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c \le j \le c, \\ j \neq 0}} \log P\left(w_{t+j}|w_t\right) \quad (1)$$

Pada persamaan (1), menunjukan fungsi objektif dari model Skip-Gram, yang bertujuan memaksimalkan probabilitas kemunculan kata-kata konteks di sekitar kata pusat. Dalam persamaan ini, T merupakan total jumlah dalam korpus, C adalah ukuran jendela konteks (window size), Wt adalah kata pusat (target word), dan $Wt = j$ adalah kata konteks (context word) yang berada di posisi ke- j dari kata pusat.
Definisi Probabilitas

$$P(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{T} v_{w_I}\right)}{\sum_{w=1}^{V} \exp\left({v'_{w_O}}^{T} v_{w_I}\right)} \quad (2)$$

Pada persamaan (2), mendefinisikan probabilitas kemunculan kata konteks $P(w_O|w_I)$ yaitu probabilitas kata konteks w_O muncul berdsarkan kata pusat w_I. Dalam persamaan ini, V adalah ukuran kosakata, V_{wI} merupakan vektor embedding dari kata pusat (vektor input), dan v'_{w0} adalah vektor embedding dari kata konteks. Fungsi exp menunjukkan operasi eksponensial.

b. GloVe
Dalam penelitian ini saya menggunakan 300 corpus GloVe yang berasal dari *api.load('glove-wiki-gigaword-300')* yang sudah dilatih oleh tim Stanford.
Rumus dasar embedding GloVe

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 \quad (3)$$

Pada persamaan (3), X_{ij} merepresentasikan jumlah kemunculan kata j di konteks kata i, sedangkan $f\left(X_{ij}\right)$ adalah fungsi pembobotan untuk mengurangi pengaruh noise dari frekuensi kata yang terlalu tinggi atau terlalu rendah. Simbol $W_i, \tilde{w}_j$ masing-masing menunjukan vektor embedding dari kata dan konteks kata, yang berasal dari dua set embedding yang berbeda. Adapun $b_i, \tilde{b}_j$ adalah nilai bias yang ditambahkan untuk kata dan konteks kata guna meningkatkan akurasi prediksi dalam model.

c. POS2Vec (P2V)
Penandaan Part-of-Speech (POS) adalah proses memberi label kelas kata seperti kata benda, kata kerja, atau kata sifat pada setiap kata dalam teks. Informasi ini membantu memahami struktur sintaksis kalimat. Dalam penelitian ini, setiap tag POS dikonversi menjadi vektor berdimensi tetap dan digabungkan dengan vektor Word2Vec atau GloVe. Penelitian ini menggunakan vektor berdimensi 50 untuk setiap kata agar mengandung informasi semantik dan sintaksis sekaligus.

d. Lexicon2Vec (L2V)
Leksikon sentimen dan emosi merupakan kumpulan kata atau frasa yang dilengkapi dengan nilai polaritas, seperti positif, negatif, atau netral, yang digunakan untuk mengukur sentimen dalam sebuah teks. Pemilihan leksikon yang tepat sangat berpengaruh terhadap akurasi analisis. Dalam penelitian ini, digunakan leksikon berbasis frekuensi kemunculan kata yang dibangun secara on-the-fly dari dataset yang digunakan. Setiap kata kemudian direpresentasikan dalam bentuk vektor berdimensi 6 berdasarkan skor dari masing-masing leksikon.

2.5 BERT

Sistem kerja algoritma BERT dimulai dengan menerima sebuah input. Input tersebut kemudian diproses oleh bagian transformer *encoder*. Di dalamnya, input akan melewati beberapa tahapan, yaitu proses token embedding, sentence embedding, dan positional embedding. Setelah proses di *encoder* selesai, hasil outputnya akan diteruskan ke bagian *decoder*. *Decoder* kemudian mengolah hasil tersebut dan mengirimkannya ke *softmax* agar matriks yang dihasilkan bisa diubah menjadi probabilitas [19].

a. Token embedding
Langkah ini melibatkan pemberian ID pada setiap kata dalam sebuah kalimat. Contohnya, kalimat "I Love BERT" diubah menjadi "[CLS] I [MASK] BERT [SEP]" setelah token-token tambahan ditambahkan. Token [CLS] digunakan untuk memberi tahu algoritma bahwa token ini berfungsi dalam prediksi kata yang sesuai atau memiliki

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

hubungan dengan kata lain dalam kalimat tersebut, sedangkan token [SEP] digunakan untuk menandai akhir dari sebuah kalimat. Contoh hasil token embedding dapat dilihat pada Tabel 1.

Tabel 1. Token embedding

I Love BERT				
$E_{[CLS]}$	E_1	$E_{[MASK]}$	E_{BERT}	$E_{[SEP]}$

b. Sentence embedding
Sentence embedding akan memberikan representasi numerik untuk membedakan antara kalimat A dan kalimat B. Contoh sentence encoding dapat dilihat pada Tabel 2.

Tabel 2. Sentence embedding

I love BERT				
E_A	E_A	E_A	E_A	E_A

c. Transformer positional encoding
Transformer positional encoding akan menempatkan penanda lokasi pada setiap kata di sebuah kalimat. Contoh Transformer positional encoding dapat dilihat pada Tabel 3.

Tabel 3. Transformer positional encoding

I love BERT				
E_0	E_1	E_2	E_3	E_4

2.6 Concatenation

Dalam pengolahan bahasa alami, concatenation adalah proses menggabungkan beberapa representasi embedding (vektor) dari kata atau frasa menjadi satu vektor tunggal. Teknik ini bertujuan untuk menyatukan informasi dari berbagai jenis atau sumber embedding, sehingga dapat memperkaya representasi semantik dan meningkatkan kinerja model pada berbagai tugas, seperti parsing, klasifikasi teks, atau pengenalan entitas. Penelitian [20] menunjukkan bahwa representasi kata yang lebih baik dapat diperoleh dengan menggabungkan beberapa jenis embedding yang saling melengkapi. Hal ini menjadi dasar diterapkannya pendekatan Automated Concatenation of Embeddings (ACE), yang digunakan dalam penelitian [20] untuk memaksimalkan potensi kombinasi embedding dalam menghasilkan model yang lebih canggih dan akurat.

$$vi = [v_i^1;\ vi_i^2, \ldots;\ v_i^L] \tag{4}$$

Pada persamaan (4) v_i merupakan representasi kata ke-i yang diperoleh melalui penggabungan beberapa vektor embedding. Simbol v_i^1 adalah embedding dari model embedding ke-l, sedangkan L menyatakan jumlah tipe embedding yang digunakan dalam proses penggabungan tersebut.
Setiap v_i^l memiliki dimensi d_i, maka dimensi total v_i adalah

$$d = \sum_{i=1}^{L} di \tag{5}$$

Pada persamaan (5), d_i adalah dimensi dari embedding ke-l. Jika terdapat L merupakan jumlah embedding yang digunakan untuk merepresentasikan kata, dan vi merupakan gabungan embedding dari L model embedding, sehingga total dimensinya adalah d.

2.7 Evaluasi

Setelah melakukan pemodelan menggunakan BERT dilakukan evaluasi model dengan menggunakan confusion matrix, *Confusion Matrix* adalah cara untuk mengevaluasi seberapa baik model *machine learning* dengan memeriksa dan membandingkan prediksi model dengan nilai yang sebenarnya. *Confusion Matrix* adalah sebuah tabel yang menampilkan empat *matrix* evaluasi yaitu TP = *true positive*, TN = *true negative*, FP = *false positive*, FN = *false negative* [1], [21]. Confusion matrix mencangkup beberapa perhitungan:

a. Akurasi
Dimana akurasi adalah nilai yang benar dibagi seluruh nilai yang ada, dan akurasi ini menggambarkan kinerja model klasifikasi dengan benar.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

d. Presisi
Presisi adalah nilai prediksi yang dimana berada didalam kelas positif dibagi dengan total nilai prediksi, pada kelas positif. Presisi menggambarkan system yang memiliki kemampuan mencari ketepatan informasi yang dimintain.

$$Presisi = \frac{TP}{TP+FP} \tag{7}$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

e. Recall
Recall dapat disebut juga sebagai rasio antara prediksi nilai benar terhadp kelas positif dengna jumalh prediksi yang benar kelas positif dan kesalahan prediksi terhadp kelas negative.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

f. F1-Score
F1-Score adalah nilai rata-rata antara presisi dan *recall,* jika nilainnya semakin mendekati angka 1 maka semakin optimal hasilnya.

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (9)$$

Keterangan: True Positive (TP) merupakan jumlah data berlabel positif yang diprediksi benar sebagai positif. True Negative (TN) adalah jumlah data berlabel negatif yang diprediksi benar sebagai negatif. False Positive (FP) menunjukan jumlah data berlabel negatif yang diprediksi salah sebagai positif, sedangkan False Negative (FN) merupakan jumlah data berlabel positif yang diprediksi salah sebagai negatif.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

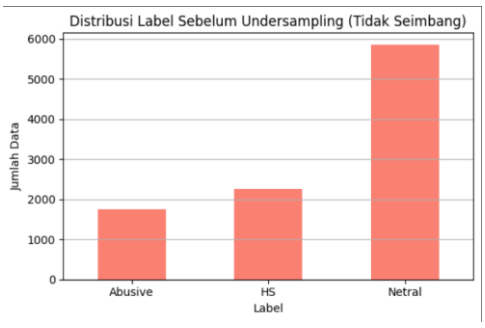
Dataset yang digunakan dalam penelitian ini merupakan hasil dari penelitia sebelumnya dan terdiri atas 13.169 data berupa unggahan (tweet) dari media sosial Twitter. Dataset ini bersifat multi-label, di mana setiap entri dapat memiliki satu atau lebih label yang merepresentasikan jenis ujaran tertentu. Label-label tersebut mengindikasikan keberadaan ujaran kebencian, kata-kata kasar, atau konten yang bersifat netral. Masing-masing tweet dilengkapi dengan atribut dalam bentuk biner, yaitu nilai 1 jika tweet mengandung jenis ujaran yang dimaksud, dan 0 jika tidak. Sepuluh contoh data awal dapat dilihat pada Gambar 2 untuk memberikan gambaran umum mengenai struktur dan isi data yang digunakan.

	Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1	0	0	0	0	0	1	1	0	0
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0	0	0	0	0	0	0	0	0	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0	0	0	0	0	0	0	0	0	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0	0	0	0	0	0	0	0	0	0	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1	0	1	1	0	0	0	0	0	1	0
5	USER Ya bani taplak dkk \xf0\x9f\x98\x84\xf0\x...	1	1	0	1	0	0	0	0	1	0	1	0
6	deklarasi pilkada 2018 aman dan anti hoax warg...	0	0	0	0	0	0	0	0	0	0	0	0
7	Gue baru aja kelar re-watch Aldnoah Zero!!! pa...	0	1	0	0	0	0	0	0	0	0	0	0
8	Nah admin belanja satu lagi port terbaik nak m...	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 2. Data Set

Dataset yang digunakan pada penelitian ini diperoleh dari penelitian [15] dan telah dilakukaan tahap preprocessing dengan total dataset sebanyak 9874. Data ini dikategorikan menjadi 3 class utama, yaitu Hate Speech, Abusive, dan Netral. Namun sebelum dilakukan proses penyeimbangan, terdapat ketidakseimbangan jumlah data antara ketiga kelas tersebut, dimana kelas Netral memiliki jumlah data yang jauh lebih besar dari pada kelas Hate Speech dan Abusive.

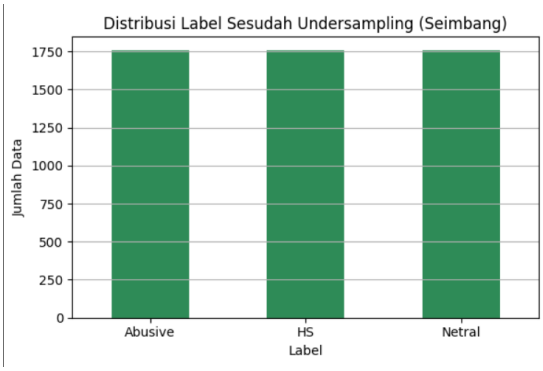
Salah satu tantangan utama yang dihadapi dalam pemrosesan dataset ini adalah ketidakseimbangan distribusi kelas. Setelah klasifikasi awal dilakukan, ditemukan bahwa jumlah data pada kelas Netral jauh lebih dominan dibandingkan dua kelas lainnya, terutama kelas Abusive yang memiliki jumlah data paling sedikit. Ketidakseimbangan ini berpotensi menimbulkan bias pada model klasifikasi, di mana model cenderung memberikan prediksi yang lebih akurat pada kelas mayoritas (Netral) dan mengabaikan kelas minoritas (Hate Speech dan Abusive). Kondisi ini tentu berdampak negatif terhadap performa keseluruhan model dalam mendeteksi konten berbahaya. Ilustrasi distribusi data antar kelas dapat dilihat pada Gambar 3.



Gambar 3. Data Tidak Seimbang

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

Untuk mengatasi permasalahan tersebut, digunakan teknik penyeimbangan data yang dikenal sebagai Random Under Sampling (RUS). Metode ini bekerja dengan cara mengurangi jumlah data pada kelas mayoritas secara acak hingga jumlahnya setara atau sebanding dengan kelas lainnya. Dengan demikian, distribusi data antar kelas menjadi lebih proporsional dan seimbang. RUS dipilih karena kesederhanaannya serta kemampuannya dalam menghilangkan dominasi kelas mayoritas tanpa memerlukan penambahan data sintetis. Meskipun metode ini memiliki risiko kehilangan informasi dari data yang dibuang, pendekatan ini tetap efektif untuk meningkatkan sensitivitas model terhadap kelas minoritas dalam konteks klasifikasi multi-kelas. Distribusi data setalah penerapan RUS dapat dilihat pada Gambar 4.



Gambar 4. Data Seimbang

3.2 Prapemrosesan

Prapemrosesan data merupakan tahap yang sangat krusial dalam mempersiapkan teks agar layak digunakan dalam proses pelatihan model klasifikasi sentimen. Tahap ini bertujuan untuk membersihkan data dari elemen-elemen yang tidak relevan, menyederhanakan struktur kalimat, serta menyelaraskan format data agar sesuai dengan kebutuhan dan standar input model berbasis BERT. Mengingat bahwa model seperti BERT sangat bergantung pada input teks yang bersih dan terstruktur untuk menghasilkan representasi yang akurat, maka proses prapemrosesan menjadi fondasi penting dalam pipeline analisis. Seluruh tahapan ini diterapkan secara konsisten pada data pelatihan maupun data pengujian guna memastikan bahwa tidak terjadi ketimpangan atau bias dalam proses pembelajaran model. Dengan kata lain, prapemrosesan tidak hanya meningkatkan kualitas data, tetapi juga berkontribusi langsung terhadap keandalan dan performa model klasifikasi sentiment.

Tahapan awal dalam proses prapemrosesan dimulai dengan tokenisasi, yaitu memecah teks atau kalimat menjadi unit-unit kecil yang disebut token, biasanya berdasarkan spasi dan tanda baca. Proses ini penting karena memungkinkan model untuk memahami struktur kalimat secara lebih rinci serta mengenali pola dalam data teks. Setelah itu dilakukan filtering, yakni menghapus kata-kata umum atau stopword yang tidak memberikan nilai informasi signifikan terhadap konteks sentimen. Pada tahap ini pula dilakukan pemilihan data berdasarkan label, yaitu HS (Hate Speech), Abusive, dan Netral.

Selanjutnya diterapkan lemmatization, yakni mengubah setiap kata ke bentuk dasarnya untuk mengurangi variasi bentuk kata yang tidak diperlukan dalam analisis. Hal ini dilanjutkan dengan proses normalisasi, yang mencakup pengubahan seluruh karakter menjadi huruf kecil (*lowercase*), serta penghapusan simbol, angka, dan karakter lain yang dianggap tidak relevan dalam konteks analisis. Perbandingan antara data sebelum dan sesudah proses *cleaning* di tampilkan pada Gambar 5.

	Tweet	cleaned_tokens
1	RT USER: USER siapa yang telat ngasih tau elu?...	[rt, user, user, telat, ngasih, tau, eluedan, ...
2	41. Kadang aku berfikir, kenapa aku tetap perc...	[kadang, berfikir, percaya, tuhan, jatuh, berk...
3	USER USER AKU ITU AKU\n\n\nKU TAU MATAMU SIPIT T...	[user, user, akunnku, tau, matamu, sipit, diliat]
6	deklarasi pilkada 2018 aman dan anti hoax warg...	[deklarasi, pilkada, aman, anti, hoax, warga, ...
7	Gue baru aja kelar re-watch Aldnoah Zero!!! pa...	[gue, aja, kelar, rewatch, aldnoah, zero, kamp...
8	Nah admin belanja satu lagi port terbaik nak m...	[admin, belanja, port, terbaik, nak, makan, ai...
9	USER Enak lg klo smbil ngewe'	[user, enak, lg, klo, smbil, ngewe]
12	Kalo belajar ekonomi mestinya jago memprivatis...	[kalo, belajar, ekonomi, mestinya, jago, mempr...
13	Aktor huruhara 98 Prabowo S ingin lengserkan p...	[aktor, huruhara, prabowo, s, lengserkan, peme...
16	Belakangan ini kok fikiran ampas banget ya'	[fikiran, ampas, banget, ya]

Gambar 5. Data Cleaning

Langkah selanjutnya adalah melakukan Part-of-Speech (POS) Tagging, yaitu proses untuk mengidentifikasi dan memberi label kategori kata, seperti kata benda, kata kerja, atau kata sifat, sesuai dengan perannya dalam kalimat. Dengan cara ini, struktur gramatikal kalimat dapat dikenali lebih baik, sehingga mendukung analisis makna teks secara lebih mendalam. Hasil dari proses POS Tagging ditunjukan pada Gambar 6, yang menampilkan setiap token beserta label kategori katanya.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

1	RT USER: USER siapa yang telat ngasih tau elu?...
2	41. Kadang aku berfikir, kenapa aku tetap perc...
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...
6	deklarasi pilkada 2018 aman dan anti hoax warg...
7	Gue baru aja kelar re-watch Aldnoah Zero!!! pa...
8	Nah admin belanja satu lagi port terbaik nak m...
9	USER Enak lg klo smbil ngewe'
12	Kalo belajar ekonomi mestinya jago memprivatis...
13	Aktor huruhara 98 Prabowo S ingin lengserkan p...
16	Belakangan ini kok fikiran ampas banget ya'

	pos_tags
1	[(RT, NNP), (USER, NNP), (:, :), (USER, NNP), ...
2	[(41, CD), (., .), (Kadang, NNP), (aku, JJ), (...
3	[(USER, NNP), (USER, NNP), (AKU, NNP), (ITU, N...
6	[(deklarasi, NN), (pilkada, NN), (2018, CD), (...
7	[(Gue, NNP), (baru, NN), (aja, NN), (kelar, JJ...
8	[(Nah, NNP), (admin, NN), (belanja, NN), (satu...
9	[(USER, NNP), (Enak, NNP), (lg, VBZ), (klo, FW...
12	[(Kalo, NNP), (belajar, NN), (ekonomi, NN), (m...
13	[(Aktor, NNP), (huruhara, VBD), (98, CD), (Pra...
16	[(Belakangan, NNP), (ini, VBZ), (kok, NN), (fi...

Gambar 6. Lexicon2Vec

Terakhir, diterapkan tahap *lexicon-based processing*, di mana digunakan basis data linguistik yang memuat kata-kata atau frasa dengan label tertentu sesuai karakteristik domain. Pendekatan ini memungkinkan sistem memiliki acuan nilai semantik tambahan untuk membantu dalam proses klasifikasi. Hasil dari tahap ini ditampilkan pada Gambar 7.

Lexicon Hate Speech:

	Word	Hate_Frequency
0	aktor	2
1	huruhara	1
2	prabowo	169
3	s	10
4	lengserkan	158
5	pemerintahan	21
6	jokowi	514
7	nyata	16
8	rt	402
9	user	2334

Lexicon Abusive Only:

	Word	Abusive_Frequency
0	rt	253
1	user	1776
2	telat	4
3	ngasih	5
4	tau	42
5	eluedan	1
6	sarap	36
7	gue	160
8	bergaul	2
9	cigax	1

Gambar 7. Pos2Vec

Seluruh tahapan prapemrosesan yang telah dilakukan mulai dari tokenisasi hingga analisis berbasis leksikon bertujuan untuk mempersiapkan teks dalam bentuk yang bersih dan terstruktur. Teks yang telah diproses tersebut kemudian diubah menjadi representasi numerik melalui teknik embedding. Pada penelitian ini, digunakan dua pendekatan untuk menghasilkan embedding, yaitu Improved Word Vector (IWV) dan BERT embedding. Kedua metode ini berfungsi untuk merepresentasikan makna kata dan konteks kalimat dalam bentuk vektor-vektor angka yang dapat dipahami oleh model klasifikasi, sehingga memungkinkan proses analisis sentimen dilakukan secara lebih akurat dan kontekstual.

3.3 Concatenation

Setelah seluruh tahapan pra-pemrosesan selesai dilakukan, langkah selanjutnya adalah membentuk representasi vektorial dari teks menggunakan teknik penggabungan embedding (concatenation). Pada tahap ini digunakan dua jenis representasi kata yang bersifat saling melengkapi, yaitu BERT embedding dan IWV. Teknik concatenation diterapkan dengan tujuan menggabungkan keunggulan masing-masing model. BERT, sebagai model berbasis arsitektur transformer, menghasilkan representasi kata yang bersifat kontekstual, yaitu mempertimbangkan makna kata berdasarkan urutan dan lingkungan katanya dalam kalimat. Kemampuan ini memungkinkan BERT menangkap konteks kalimat secara lebih mendalam dan dinamis. Sementara itu, IWV merupakan representasi kata yang diperoleh dari model pembelajaran tradisional seperti Word2Vec atau GloVe, yang memanfaatkan kemunculan kata dalam korpus besar untuk membentuk makna secara statistik. Representasi ini bersifat lebih statis namun tetap memberikan informasi semantik yang penting. Dengan menggabungkan embedding dari BERT dan IWV, diharapkan representasi teks yang dihasilkan dapat memuat informasi yang lebih lengkap baik dari segi konteks maupun makna sehingga mendukung peningkatan akurasi dalam proses klasifikasi sentiment.

3.4 Klasifikasi BERT

Setelah tahap pra-pemrosesan selesai dilakukan dan embedding dari IWV serta BERT digabungkan, tahap selanjutnya adalah melakukan pelatihan model menggunakan arsitektur BERT yang diperkuat dengan tambahan embedding dari IWV. BERT merupakan pendekatan berbasis transformer yang dirancang khusus untuk memahami konteks dalam pemrosesan bahasa alami. Dalam penelitian ini, BERT digunakan sebagai model utama untuk mengklasifikasikan teks

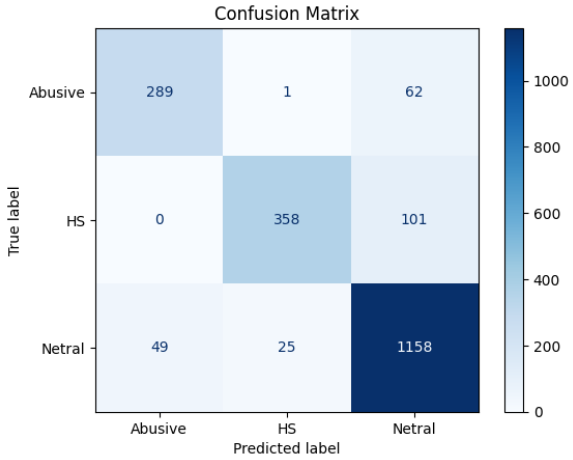
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

ke dalam tiga kategori, yaitu Hate Speech, Abusive, dan Netral. Arsitektur BERT bekerja dengan memanfaatkan mekanisme self-attention secara dua arah, sehingga mampu memahami konteks kata dalam kalimat baik dari arah kiri maupun kanan. Setiap kata atau token akan diproses melalui beberapa lapisan encoder transformer, dan token khusus [CLS] digunakan sebagai representasi keseluruhan teks input. Embedding gabungan dari IWV dan BERT digunakan sebagai masukan ke dalam model, sehingga memperkaya informasi makna dan kontekstual yang diterima oleh sistem. Proses pelatihan dilakukan dengan menggunakan fungsi loss yang sesuai (seperti cross-entropy loss), dan optimisasi dilakukan untuk meminimalkan kesalahan prediksi terhadap label sebenarnya. Output dari proses ini adalah prediksi label klasifikasi untuk setiap teks, yang menunjukkan apakah suatu teks tergolong Hate Speech, Abusive, atau Netral, beserta probabilitasnya. Selain itu, evaluasi dilakukan menggunakan metrik seperti precision, recall, f1-score, dan akurasi untuk mengukur performa model dalam tugas klasifikasi.

3.5 Evaluasi

Evaluasi dilakukan untuk mengukur performa model dalam melakukan klasifikasi sentimen berdasarkan representasi teks yang dihasilkan dari gabungan IWV dan BERT embedding. Evaluasi dilakukan dalam dua skenario berbeda, yaitu sebelum dan sesudah dilakukan penyeimbangan data menggunakan teknik Random Under Sampling (RUS). Metrik evaluasi yang digunakan mencakup akurasi, presisi, recall, dan f1-score yang dihitung berdasarkan nilai per kelas maupun nilai rata-rata (macro dan weighted average).

a. Model IWV+BERT (Tanpa Penyeimbang Data)
Model pertama dilatih menggunakan data asli yang belum melalui proses penyeimbangan kelas. Tujuan dari skenario ini adalah untuk mengevaluasi sejauh mana model dapat mengklasifikasikan sentimen tanpa intervensi terhadap distribusi kelas. Dataset yang digunakan masih bersifat tidak seimbang, di mana jumlah data kelas Netral jauh lebih besar dibandingkan kelas Abusive dan Hate Speech. Hal ini dapat memberikan pengaruh terhadap hasil prediksi model, khususnya menimbulkan bias terhadap kelas mayoritas. Hasil dari pengujiian ini ditampilkan pada Gambar 8 dalam bentuk *confusion matrix*.



Gambar 8. Hasil IWV+BERT (Tanpa Penyeimbang Data)

Tabel 4 berikut menunjukkan hasil evaluasi model pada data yang tidak seimbang:

Tabel 4. Hasil IWV+BERT (Tanpa Penyeimbang Data)

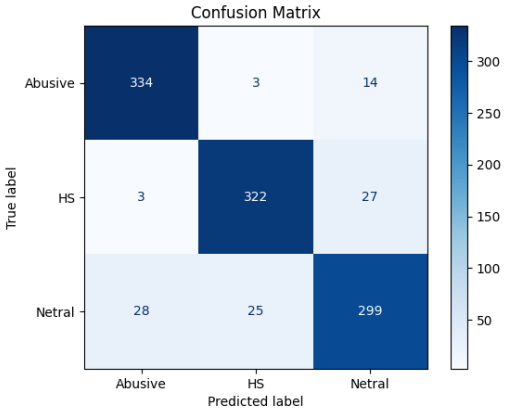
	precision	recall	f1-score	Support
Abusive	0.86	0.82	0.84	352
Hate Speech	0.93	0.78	0.85	459
Netral	0.88	0.94	0.91	1232
acuracy			0.88	2043
macro avg	0.89	0.89	0.86	2043
weighted avg	0.89	0.88	0.88	2043

Dari hasil evaluasi pada Tabel 4, terlihat bahwa model memiliki performa yang cukup baik secara umum, dengan akurasi sebesar 88%. Namun, terdapat perbedaan performa antar kelas, khususnya pada kelas Hate Speech yang meskipun memiliki presisi tinggi (0.93), namun nilai recall-nya lebih rendah (0.78). Hal ini menunjukkan bahwa model cenderung kurang mampu menangkap semua data yang benar-benar termasuk dalam kelas tersebut, yang kemungkinan disebabkan oleh ketidakseimbangan data.

b. Model IWV+BERT (Dengan Penyeimbang Data)
Model kedua dilatih menggunakan data yang telah diseimbangkan menggunakan teknik Random Under Sampling (RUS). Tujuan dari skenario ini adalah untuk mengetahui dampak penyeimbangan distribusi kelas terhadap

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

performa model. RUS dilakukan dengan cara mengurangi jumlah data pada kelas mayoritas (Netral) secara acak, sehingga jumlah data pada ketiga kelas menjadi seimbang. Dengan demikian, model diharapkan dapat belajar secara lebih adil terhadap masing-masing kelas. Hasil dari pengujiian ini ditampilkan pada Gambar 9 dalam bentuk *confusion matrix*.



Gambar 9. Hasil IWV+BERT (Dengan Penyeimbang Data)

Tabel 5 berikut menampilkan hasil evaluasi model pada data yang telah seimbang:

Tabel 5. Hasil IWV+BERT (Dengan Penyeimbang Data)

	precision	recall	f1-score	Support
Abusive	0.92	0.95	0.93	351
Hate Speech	0.92	0.91	0.92	352
Netral	0.88	0.85	0.86	352
acuracy			0.91	1055
macro avg	0.90	0.91	0.90	1055
weighted avg	0.90	0.91	0.90	1055

Dari hasil evaluasi pada Tabel 5, terlihat bahwa performa model mengalami peningkatan secara menyeluruh. Akurasi meningkat menjadi 91%, dan ketiga kelas memperoleh nilai presisi, recall, dan f1-score yang relatif seimbang. Hal ini menunjukkan bahwa proses penyeimbangan data mampu meningkatkan kemampuan model dalam mengenali masing-masing kelas dengan lebih baik, khususnya kelas minoritas seperti Abusive dan Hate Speech yang sebelumnya cenderung terabaikan.

3.6 Pembahasan

Berdasarkan hasil evaluasi model BERT yang diperkuat dengan tambahan embedding IWV dalam dua eksperimen yang telah dilakukan pada penelitian ini, dapat disimpulkan bahwa keberhasilan model klasifikasi dengan data seimbang menghasilkan akurasi yang lebih baik dibandingkan dengan data yang tidak seimbang. Tahap pra-pemrosesan pada dataset sangat berpengaruh terhadap hasil akhir. Penggunaan teknik penyeimbangan data terbukti menjadi faktor penting dalam peningkatan performa model. Teknik ini berperan krusial dalam meningkatkan keadilan klasifikasi, terutama untuk kelas minoritas. Tanpa penyeimbangan, model cenderung mengabaikan kelas minoritas akibat dominasi kelas mayoritas pada data asli.

Pada kondisi awal, ketika model BERT dilatih menggunakan data yang tidak seimbang, performa model sudah menunjukkan hasil yang cukup baik secara metrik akurasi yaitu sebesar 81%, namun masih menunjukkan kelemahan dalam mendeteksi kelas-kelas minoritas seperti Hate Speech dan Netral. Hal ini tampak dari nilai F1-score pada kelas Hate Speech hanya sebesar 0,73 dan kelas Netral sebesar 0,82, lebih rendah dibandingkan dengan kelas mayoritas Abusive yang mencapai 0,87. Model menunjukkan bias terhadap kelas mayoritas, sehingga prediksi terhadap kasus cyberbullying yang sebenarnya masih berpotensi terlewat

Setelah dilakukan penyeimbangan data menggunakan metode Random Under Sampling (RUS) dan integrasi embedding IWV, model menunjukkan peningkatan performa secara menyeluruh. Model BERT+IWV menghasilkan akurasi sebesar 91%, meningkat 10% dibandingkan model BERT sebelumnya yang hanya mencapai akurasi 81% [1]. Selain itu, peningkatan metrik juga tercermin pada kelas minoritas. Misalnya, F1-score untuk kelas Hate Speech meningkat dari 73% menjadi 92%, sedangkan untuk kelas Netral meningkat dari 82% menjadi 86%. Kelas Abusive juga menunjukkan peningkatan menjadi 93%.

Peningkatan ini menunjukkan bahwa model tidak hanya lebih akurat secara umum, tetapi juga lebih seimbang dalam mendeteksi setiap kategori, termasuk kelas minoritas yang sebelumnya sulit dikenali. Secara praktis, hal ini sangat penting dalam konteks deteksi cyberbullying di dunia nyata, karena model sekarang mampu mengenali lebih banyak kasus Hate Speech atau Abusive Language dengan recall tinggi (91% dan 95%) tanpa mengorbankan precision

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

(92% dan 92%). Artinya, model mampu mendeteksi lebih banyak kasus sebenarnya (true positive) tanpa terlalu banyak menghasilkan false positive, yang merupakan tantangan utama dalam sistem pendeteksi otomatis.

Namun demikian, penelitian ini memiliki keterbatasan pada aspek generalisasi, mengingat seluruh proses pelatihan dan evaluasi hanya menggunakan satu jenis dataset, yaitu dataset yang digunakan dalam penelitian [15]. Keterbatasan ini dapat memengaruhi kemampuan model dalam menangani data dari domain lain atau dengan karakteristik distribusi yang berbeda. Oleh karena itu, untuk penelitian selanjutnya disarankan agar metode ini diuji pada berbagai dataset dari sumber berbeda, termasuk yang memiliki variasi bahasa, konteks sosial, atau struktur ujaran yang lebih kompleks. Selain itu, eksplorasi terhadap metode penyeimbangan data yang lebih canggih, seperti SMOTE atau Generative Data Augmentation, serta integrasi dengan model transformer-based lainnya dapat menjadi arah pengembangan lebih lanjut dalam meningkatkan ketahanan dan akurasi sistem deteksi cyberbullying secara umum.

Penerapan fitur ekstraksi Improved Word Vector (IWV) pada data teks sentimen terkait cyberbullying di media sosial Twitter (X) dilakukan melalui serangkaian tahapan preprocessing yang sistematis dan terstruktur. Tahapan ini mencakup beberapa langkah penting, dimulai dari tokenisasi, yaitu proses memecah kalimat menjadi unit-unit kata. Selanjutnya dilakukan filtering terhadap kata-kata tidak penting menggunakan daftar stopword, serta proses stemming atau lemmatization untuk mereduksi kata ke bentuk dasarnya guna mengurangi variasi morfologis. Selain itu, dilakukan pula normalisasi teks untuk memastikan konsistensi format, seperti mengubah seluruh huruf menjadi lowercase, menghapus tanda baca, angka, serta simbol-simbol yang tidak relevan dengan konteks analisis.

Dalam proses pembentukan IWV, ditambahkan pula dua fitur linguistik tambahan yang memperkaya representasi vektor, yaitu POS2Vec, yang berbasis Part-of-Speech tagging, dan Lexicon2Vec, yang memanfaatkan sumber data leksikal untuk mengekstraksi nilai sentimen dari kata-kata berdasarkan domain tertentu. Hasil akhir dari proses ini adalah vektor kata berdimensi 665, yang mencerminkan kombinasi dari berbagai sumber informasi linguistik dan semantik. Untuk meningkatkan performa model, dilakukan penggabungan antara IWV dan embedding BERT melalui proses concatenation pada level token maupun kalimat, sehingga diperoleh representasi akhir berdimensi 1424. Gabungan ini terbukti lebih efektif dibandingkan penggunaan BERT saja, terutama pada data yang telah melalui proses penyeimbangan kelas, dengan akurasi tertinggi yang berhasil dicapai sebesar 91%.

4. KESIMPULAN

Penerapan fitur ekstraksi Improved Word Vector (IWV) pada data teks sentimen terkait cyberbullying di media sosial Twitter (X) dilakukan melalui serangkaian tahapan preprocessing yang sistematis dan terstruktur. Tahapan ini mencakup beberapa langkah penting, dimulai dari tokenisasi, yaitu proses memecah kalimat menjadi unit-unit kata. Selanjutnya dilakukan filtering terhadap kata-kata tidak penting menggunakan daftar stopword, serta proses stemming atau lemmatization untuk mereduksi kata ke bentuk dasarnya guna mengurangi variasi morfologis. Selain itu, dilakukan pula normalisasi teks untuk memastikan konsistensi format, seperti mengubah seluruh huruf menjadi lowercase, menghapus tanda baca, angka, serta simbol-simbol yang tidak relevan dengan konteks analisis. Dalam proses pembentukan IWV, ditambahkan pula dua fitur linguistik tambahan yang memperkaya representasi vektor, yaitu POS2Vec, yang berbasis Part-of-Speech tagging, dan Lexicon2Vec, yang memanfaatkan sumber data leksikal untuk mengekstraksi nilai sentimen dari kata-kata berdasarkan domain tertentu. Hasil akhir dari proses ini adalah vektor kata berdimensi 665, yang mencerminkan kombinasi dari berbagai sumber informasi linguistik dan semantik. Untuk meningkatkan performa model, dilakukan penggabungan antara IWV dan embedding BERT melalui proses concatenation pada level token maupun kalimat, sehingga diperoleh representasi akhir berdimensi 1424. Gabungan ini terbukti lebih efektif dibandingkan penggunaan BERT saja, terutama pada data yang telah melalui proses penyeimbangan kelas, dengan akurasi tertinggi yang berhasil dicapai sebesar 91%.

REFERENCES

[1] N. Putu, V. D. Saraswati, N. Yudistira, and P. P. Adikara, "Analisis Sentimen terhadap Perundungan Siber pada Twitter menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, pp. 909–916, 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12345

[2] C. Destitus, W. Wella, and S. Suryasari, "Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia," *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 11, no. 2, pp. 107–111, 2020, doi: 10.31937/si.v11i2.1740.

[3] Z. Mansur, N. Omar, and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities," *IEEE Access*, vol. 11, no. February, pp. 16226–16249, 2023, doi: 10.1109/ACCESS.2023.3239375.

[4] A. R. Hakim, D. M. U. Atmaja, D. Haryadi, and N. Suwaryo, "Tik-53 Twitter Sentiment Analysis Terhadap Pengguna E-Commerce Menggunakan Text Mining," *Pros. Semin. Nas. Teknol. Energi dan Miner.*, vol. 1, no. 2, pp. 1227–1237, 2021, doi: 10.53026/sntem.v1i2.592.

[5] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine," *JST (Jurnal Sains dan Teknol.*, vol. 12, no. 1, pp. 129–137, 2023, doi: 10.23887/jstundiksha.v12i1.52358.

[6] D. Nugraha and P. Astuti, "Analisis Sentimen Cyberbullying Pada Sosial Media Instagram Menggunakan Metode Support Vector Machine," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 8, no. 2, p. 153, 2023, doi: 10.51211/isbi.v8i2.2535.

[7] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1057−1068
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7968

Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[8] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16–22, 2021, doi: 10.30813/jbase.v4i2.3000.

[9] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.

[10] M. Khadapi and V. Maruli Pakpahan, "Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024," *JUKI J. Komput. dan Inform.*, vol. 6, no. 2, pp. 130–137, 2024, [Online]. Available: https://www.ioinformatic.org/index.php/JUKI/article/view/681

[11] M. Adrinta Abdurrazzaq and E. Lesmana Tjiong, "Analisis Sentimen KUHP Baru Pada Data Twitter Menggunakan BERT," *J. Komunikasi, Sains dan Teknol.*, vol. 1, no. 2, pp. 83–88, 2022, doi: 10.61098/jkst.v1i2.10.

[12] C. A. Putri, "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 2, pp. 181–193, 2020, doi: 10.35957/jatisi.v6i2.206.

[13] B. Kurniawan, A. Ari Aldino, and A. Rahman Isnain, "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (PSE) Menggunakan Algoritma Bidirectional Encorder Representations From Transformer (BERT)," *J. Teknol. dan Sist. Inf.*, vol. 3, no. 4, pp. 98–106, 2022, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[14] S. M. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis," *arXiv*, 2017, [Online]. Available: http://arxiv.org/abs/1711.08609

[15] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *Proc. Third Work. Abus. Lang. Online*, pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.

[16] S. Riadi, E. Utami, and A. Yaqin, "Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection," *Sinkron*, vol. 8, no. 4, pp. 2414–2424, 2023, doi: 10.33395/sinkron.v8i4.12629.

[17] M. H. Fariz and E. B. Setiawan, "the Impact of Word Embedding on Cyberbullying Detection Using Hybird Deep Learning Cnn-Bilstm," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 10, no. 3, pp. 661–671, 2025, doi: 10.33480/jitk.v10i3.6270.

[18] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.

[19] A. Nayla, C. Setianingsih, and B. Dirgantoro, "Deteksi Hate Speech Pada Twitter Menggunakan Algoritma BERT," *e-Proceeding Eng.*, vol. 10, no. 1, p. 256, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19323

[20] X. Wang, Z. Jia, Y. Jiang, and K. Tu, "Enhanced Universal Dependency Parsing with Automated Concatenation of Embeddings," *arXiv*, pp. 189–195, 2021, doi: 10.18653/v1/2021.iwpt-1.20.

[21] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.