JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.7176

ANALISIS SENTIMEN PUBLIK PADA TAGAR #BTSCOMEBACK DI PLATFORM X MENGGUNAKAN INDOBERTWEET

Nastasya Meryl Damayanti[1], Imelda Dwi Ariningtyas[2], Maulana Izuddin Audadi Icham[3], Anggraini Puspita Sari[4*]

[1234]Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Jalan Raya Rungkut Madya No. 1, Gunung Anyar, Surabaya, Indonesia.

Keywords:
Analisis Sentimen;
IndoBERTweet;
BTS *comeback;*
Media Sosial.

Corespondent Email:
anggraini.puspita.if@upnjatim.ac.id

Abstrak. Media sosial telah menjadi ruang utama bagi publik dalam mengekspresikan opini terhadap fenomena budaya populer, termasuk *comeback* grup K-pop BTS yang yang sering kali menimbulkan intensitas percakapan dan partisipasi digital. Tagar #BTSComeback menjadi salah satu kanal ekspresi publik yang ramai digunakan, mencerminkan beragam respons dari pengguna internet di Indonesia, mulai dari dukungan antusias hingga bentuk kritik. Penelitian ini bertujuan untuk menganalisis sentimen publik Indonesia terhadap tagar tersebut dengan memanfaatkan model IndoBERTweet, yaitu model pralatih yang dirancang khusus untuk memahami teks berbahasa Indonesia di media sosial. Sebanyak 6.300 tweet berbahasa Indonesia dikumpulkan dari platform X dalam rentang waktu Januari hingga Juni 2025. Hasil penelitian menunjukkan bahwa IndoBERTweet mampu mengklasifikasikan sentimen dengan akurasi mencapai 95%, serta menghasilkan performa evaluasi yang konsisten tinggi pada ketiga kategori sentimen, terutama dalam mendeteksi sentimen positif. Visualisasi dalam bentuk *word cloud* memperlihatkan keberagaman ekspresi publik terhadap peristiwa *comeback* tersebut. Penelitian ini membuktikan efektivitas IndoBERTweet dalam menganalisis sentiment teks media sosial berbahasa Indonesia dan memberikan wawasan empiris tentang dinamika opini publik Indonesia terhadap fenomena budaya popular global.

Abstract. *Social media has become the main space for the public to express opinions on popular culture phenomena, including the comeback of K-pop group BTS which often generates intense conversations and digital participation. The hashtag #BTSComeback has become one of the most widely used channels of public expression, reflecting various responses from internet users in Indonesia, ranging from enthusiastic support to criticism. This study aims to analyze Indonesian public sentiment towards the hashtag by utilizing the IndoBERTweet model, a pre-trained model specifically designed to understand Indonesian-language texts on social media. A total of 6,300 Indonesian-language tweets were collected from platform X between January and June 2025. The results showed that IndoBERTweet was able to classify sentiment with an accuracy of up to 95%, and produced consistently high evaluation performance in all three sentiment categories, especially in detecting positive sentiment. Visualization in the form of a word cloud shows the diversity of public expression towards the comeback event. This study proves the effectiveness of IndoBERTweet in analyzing sentiment on Indonesian-language social media texts and provides empirical insight into*

the dynamics of Indonesian public opinion towards global popular culture phenomena.

1. PENDAHULUAN

Dalam beberapa tahun terakhir, perkembangan teknologi komunikasi telah memberikan dampak yang signifikan[1]. Perkembangan teknologi komunikasi telah mendorong transformasi cara masyarakat menyampaikan opini, salah satunya melalui platform digital seperti X (sebelumnya Twitter) [2]. Platform ini menjadi ruang diskusi yang aktif dan berpengaruh dalam membentuk opini publik berskala global. Dalam konteks budaya populer, gelombang K-pop telah menjadi katalis munculnya dinamika komunikasi digital yang intens, cepat, dan masif. BTS (Bangtan Sonyeondan), sebagai salah satu ikon terbesar dari industri musik Korea Selatan, tidak hanya mendominasi panggung hiburan internasional, tetapi juga membentuk pola interaksi digital di berbagai belahan dunia termasuk di Indonesia melalui pengaruh yang kuat dalam komunitas daring mereka. Ketika BTS mengumumkan comeback atau merilis karya baru, tagar seperti #BTSComeback secara konsisten mendominasi *trending topic global* di platform X dalam hitungan menit. Fenomena ini mencerminkan kekuatan media sosial sebagai ruang ekspresi emosional kolektif, di mana jutaan pengguna dari berbagai negara, termasuk Indonesia, menyampaikan antusiasme, ekspektasi, kritik, atau berbagai bentuk respons emosional lainnya. Intensitas dan volume interaksi yang terjadi menciptakan kumpulan data opini publik yang sangat kaya dan beragam, sehingga menjadi objek penelitian yang menarik untuk dianalisis menggunakan pendekatan analisis sentimen [3].

Analisis sentimen merupakan teknik pemrosesan bahasa alami yang bertujuan untuk mengidentifikasi dan mengekstrak informasi subjektif dari teks, khususnya untuk menentukan polaritas emosi atau opini yang terkandung dalam suatu dokumen atau kalimat[4]. Dalam konteks media sosial, analisis sentimen memiliki peran strategis dalam memahami persepsi publik terhadap berbagai isu, brand, produk, atau fenomena budaya. Kemampuan untuk mengklasifikasikan opini menjadi kategori positif, negatif, atau netral memberikan wawasan mendalam tentang dinamika emosi kolektif dan tren sosial yang berkembang di masyarakat digital [5]. Penerapan analisis sentimen pada teks media sosial, khususnya Twitter, menghadapi tantangan teknis yang signifikan. Karakteristik unik dari teks Twitter yang informal, penggunaan slang, emotikon, hashtag, *mention*, singkatan, dan struktur bahasa yang tidak baku membuat model analisis sentimen konvensional sering mengalami penurunan performa. Untuk mengatasi tantangan ini, diperlukan model yang secara khusus dirancang dan dilatih untuk memahami konteks dan struktur bahasa informal media sosial[6]. IndoBERTweet merupakan model transformasi dari BERT (*Bidirectional Encoder Representations from Transformers*) yang telah diadaptasi khusus untuk memproses teks bahasa Indonesia di platform Twitter [7]. Model ini menggunakan pendekatan *domain-specific vocabulary adaptation*, yang menambahkan kosakata khas Twitter seperti slang, hashtag, emotikon, dan mention ke dalam model IndoBERT tanpa melakukan pretraining dari nol[8]. Keunggulan IndoBERTweet terletak pada kemampuannya mengenali dan memproses struktur bahasa informal yang khas media sosial Indonesia, sehingga lebih efektif dalam menganalisis sentimen dibandingkan model BERT standar[9]. Penelitian sebelumnya telah menunjukkan efektivitas model BERT dan variannya dalam berbagai tugas analisis sentimen. Namun, belum banyak penelitian yang secara spesifik mengeksplorasi penerapan IndoBERTweet untuk menganalisis sentimen publik terhadap fenomena budaya populer seperti *comeback* grup K-pop di Indonesia[10]. Sebagian besar penelitian analisis sentimen di Indonesia masih berfokus pada topik politik, ekonomi, atau isu sosial secara umum, sementara kajian terhadap respons emosional publik Indonesia terhadap peristiwa budaya populer global yang memiliki dampak luas masih tergolong terbatas[11].

Pemanfaatan model IndoBERTweet dalam penelitian ini didasarkan pada kemampuannya dalam memahami ragam bahasa Indonesia yang

digunakan di media sosial, termasuk bentuk-bentuk informal, singkatan, dan struktur kalimat yang tidak baku yang umum dijumpai dalam tweet. Tagar #BTSComeback dipilih sebagai studi kasus karena mewakili fenomena budaya populer global yang mendapat perhatian luas dari publik Indonesia, serta menjadi topik yang aktif dibicarakan di platform X saat ini. Kombinasi antara studi kasus yang relevan secara kontekstual dan pemilihan pendekatan pemodelan bahasa yang tepat menjadikan metode ini efektif dalam mengungkap sentimen publik secara lebih akurat dan representatif, khususnya dalam konteks interaksi dan ekspresi opini masyarakat Indonesia di media sosial. Analisis sentimen terhadap fenomena ini tidak hanya memberikan wawasan tentang dinamika fandom, tetapi juga dapat memberikan pemahaman tentang bagaimana masyarakat Indonesia merespons dan mengadopsi budaya populer global[13]. Selain itu, penelitian ini juga memiliki nilai praktis dalam konteks pemahaman perilaku konsumen digital dan strategi komunikasi pemasaran. Hasil analisis sentimen dapat memberikan insight berharga bagi industri hiburan, manajemen artis, dan platform media sosial dalam memahami preferensi dan respons emosional audiens Indonesia terhadap konten K-pop. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metodologi analisis sentimen untuk teks media sosial berbahasa Indonesia, khususnya dalam konteks budaya populer. Selain itu, hasil penelitian ini juga diharapkan dapat memberikan wawasan empiris tentang dinamika opini publik Indonesia terhadap fenomena budaya global, yang dapat bermanfaat bagi penelitian selanjutnya di bidang komunikasi digital, studi budaya populer, dan analisis media sosial. Gap penelitian ini menjadi semakin relevan mengingat besarnya pengaruh industri K-pop, khususnya BTS, terhadap masyarakat Indonesia. Data menunjukkan bahwa Indonesia merupakan salah satu negara dengan jumlah penggemar K-pop terbesar di dunia, dan setiap aktivitas BTS selalu memicu respons masif di media sosial Indonesia[12].

2. TINJAUAN PUSTAKA

1.1. Analisis Sentimen

Analisis sentimen merupakan proses otomatis dalam mengolah dan memahami data teks yang tidak terstruktur guna mengidentifikasi serta memperoleh informasi mengenai kecenderungan opini atau sikap yang terkandung dalam sebuah pernyataan atau ungkapan[14]. Analisis sentimen berperan dalam mengungkap sikap, pandangan, serta emosi yang tersirat dalam sebuah teks, sehingga dapat memberikan gambaran yang lebih komprehensif terhadap respons pengguna terhadap suatu isu, baik dalam bentuk dukungan, penolakan, maupun ketidakpastian, yang pada akhirnya dapat digunakan untuk memahami kecenderungan opini publik secara lebih sistematis dan terstruktur[15]. Tujuan utama dari analisis sentimen adalah untuk mengidentifikasi dan mengklasifikasikan kecenderungan opini yang terdapat dalam sebuah teks, baik secara eksplisit maupun implisit, guna menentukan apakah sentimen tersebut bersifat positif, negatif, atau netral berdasarkan konteks, struktur kalimat, dan pemilihan kata yang digunakan oleh penulis dalam menyampaikan pandangannya[16].

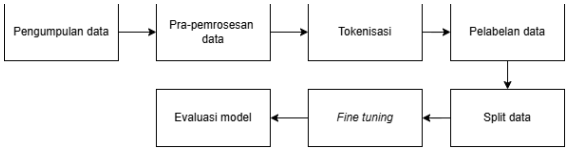
2.2. IndoBERTweet

BERT merupakan model representasi bahasa pra-terlatih yang menggunakan arsitektur Transformer dengan mekanisme *self-attention.* Model ini memanfaatkan arsitektur Transformer yang mampu membaca konteks kata dari dua arah (kiri dan kanan), sehingga menghasilkan embedding kontekstual yang sangat kaya dan unggul dalam berbagai tugas seperti klasifikasi sentimen, parsing, dan pemahaman teks secara umum[17]. Namun, ketika BERT diterapkan pada teks bahasa Indonesia, muncul kebutuhan untuk mengadaptasi model ini agar lebih sesuai dengan karakteristik bahasa dan konteks lokal. Dari sini, lahirlah IndoBERT, yang merupakan versi BERT yang telah disesuaikan untuk bahasa Indonesia[18]. IndoBERT mempertahankan keunggulan arsitektur BERT, tetapi dengan kosakata dan struktur yang lebih relevan untuk pengguna bahasa Indonesia. Untuk lebih meningkatkan performa dalam analisis sentimen di platform media sosial seperti Twitter, dikembangkanlah IndoBERTweet. Model ini tidak hanya menggunakan IndoBERT sebagai dasar, tetapi juga melakukan *domain-specific vocabulary adaptation* dengan menambahkan kosakata khas Twitter, seperti slang, hashtag, emotikon,

dan mention ke dalam IndoBERT tanpa melakukan pretraining dari nol[19]. Embedding untuk token baru ini diinisialisasi menggunakan rata-rata representasi subtoken dari model BERT asli, menghasilkan peningkatan efisiensi hingga lima kali lebih cepat dibandingkan pretraining penuh, sekaligus lebih efektif secara performa ekstrinsik [20]. Model ini dibangun dengan 128 layer dan memanfaatkan fungsi aktivasi sigmoid. Proses tokenisasi dilakukan menggunakan AutoTokenizer dengan model dasar indolem. Dalam proses ini, token khusus seperti [CLS] disisipkan untuk menyesuaikan struktur input, mengikuti pendekatan yang serupa dengan arsitektur BERT[21].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan pemodelan bahasa berbasis IndoBERTweet untuk menganalisis sentimen publik terhadap tagar #BTScomeback di platform X. Metode yang digunakan terdiri dari beberapa tahapan utama, yaitu pengumpulan data, pra-pemrosesan data, tokenisasi, pelabelan data, split data, *fine-tuning*, dan evaluasi model sebagaimana ditunjukkan apada Gambar 1.



Gambar 1. Diagram alur metodologi

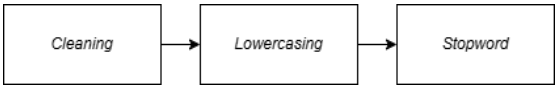
3.1. Pengumpulan Data

Data dalam penelitian ini diperoleh dari platform X dengan menggunakan tagar #BTScomeback sebagai kata kunci utama dalam proses pengambilan data. Pengumpulan data dilakukan secara otomatis melalui teknik crawling dengan memanfaatkan *authentication token* (*auth token*) untuk memperoleh akses yang sah dan terautentikasi terhadap API Twitter. Proses ini memudahkan dalam mengumpulkan tweet dalam jumlah besar secara otomatis, tanpa harus mengambil data secara manual. Selain itu, penggunaan auth token juga memastikan bahwa proses pengambilan data berlangsung secara legal dan sesuai prosedur teknis dari platform.

Rentang waktu pengambilan data ditetapkan mulai dari Januari hingga Juni 2025, yang bertepatan dengan periode comeback BTS. Penetapan rentang waktu ini bertujuan untuk menangkap dinamika opini publik secara langsung selama momen tersebut berlangsung. Dengan membatasi pengambilan data pada periode tersebut, hasil analisis dapat mencerminkan persepsi publik yang aktual dan relevan terhadap topik yang dibahas. Hasil dari proses crawling menghasilkan sebanyak 6.300 tweet yang berkaitan dengan tagar #BTScomeback. Seluruh data yang diperoleh disimpan dalam format file .csv, sehingga memudahkan dalam tahap pra-pemrosesan dan analisis lebih lanjut. Dataset ini terdiri atas teks asli dari tweet yang akan dianalisis untuk mengidentifikasi kecenderungan sentimen publik terhadap topik yang dibahas.

3.2. Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap penting dalam pengolahan data teks, yang bertujuan untuk meningkatkan kualitas data sebelum dianalisis lebih lanjut. Proses ini dilakukan untuk menghilangkan elemen-elemen yang tidak diperlukan, menyederhanakan struktur teks, serta menyiapkan data agar sesuai dengan format yang dibutuhkan oleh model analisis sentimen. Tahapan pra-pemrosesan dalam penelitian ini mencakup proses *cleaning*, *lowercasing*, dan penghapusan *stopword*, sebagaimana terlihat pada Gambar 2.



Gambar 2. Tahap pra-pemrosesan data

3.2.1 *Cleaning*

Cleaning text merupakan tahap awal dalam pra-pemrosesan data yang bertujuan untuk membersihkan teks dari elemen-elemen yang tidak relevan atau mengganggu analisis. Pada tahap ini, dilakukan penghapusan berbagai karakter seperti tautan (URL), mention (@username), hashtag (#), angka, tanda baca yang tidak diperlukan, emotikon, serta karakter non-alfabet lainnya. Proses ini penting untuk memastikan bahwa teks yang dianalisis memiliki struktur yang lebih rapi dan fokus pada isi utama dari tweet, sehingga dapat

meningkatkan akurasi dalam tahap klasifikasi sentimen yang terlihat pada Tabel 1.

Tabel 1. *Cleaning Text*

Sebelum	Sesudah
gak sabar bts comeback >__<	gak sabar bts comeback gt__lt

3.2.2 *Lowercasing*

Lowercasing merupakan proses transformasi seluruh karakter dalam teks menjadi huruf kecil dengan tujuan untuk menyamakan representasi kata secara konsisten. Langkah ini penting agar sistem tidak membedakan kata yang sama akibat perbedaan penggunaan huruf kapital, seperti yang terletak pada Tabel 2, kata "BTS" dan "bts" yang seharusnya dikenali sebagai satu entitas. Dengan menerapkan lowercasing, proses analisis teks menjadi lebih terstruktur dan efektif, serta mengurangi duplikasi kata dalam proses tokenisasi dan klasifikasi sentimen.

Tabel 2. Proses *Lowercasing*

Sebelum	Sesudah
gak sabar bts comeback >_<	gak sabar bts comeback gt__lt

3.2.3 *Stopword*

Stopword merupakan kata-kata umum yang sering muncul dalam teks namun tidak memiliki makna signifikan dalam proses analisis, seperti "yang", "dan", "di", atau "untuk". Pada Tabel 3, proses penghapusan stopword dilakukan guna mengurangi noise dalam data dan memfokuskan analisis pada kata-kata yang memiliki bobot informasi yang lebih kuat. Dengan mengeliminasi *stopword*, representasi teks menjadi lebih ringkas dan relevan, sehingga dapat meningkatkan performa model dalam mengklasifikasikan sentimen dari setiap tweet.

Tabel 3. Proses *Stopword*

Sebelum	Sesudah
gak sabar bts comeback >__<	gak sabar bts comeback gt__l

3.3. Tokenisasi

Tokenisasi merupakan proses awal yang bertujuan untuk memecah teks mentah menjadi unit-unit terkecil yang disebut token, seperti kata, sub-kata, atau karakter. Proses ini memungkinkan teks diubah ke dalam bentuk representasi numerik yang dapat diproses oleh model pembelajaran mesin. Dalam penelitian ini, tokenisasi dilakukan terhadap teks tweet berbahasa Indonesia yang telah melalui tahap pembersihan (*cleaning*). Tokenisasi menggunakan *AutoTokenizer* dari pustaka *transformers* milik *Hugging Face* dengan basis model *indolem/indobertweet-base-uncased*, yang merupakan tokenizer resmi untuk model IndoBERTweet. Tokenizer ini mampu mengenali struktur bahasa informal khas media sosial, seperti singkatan, emotikon, atau kata tidak baku dalam bahasa Indonesia. Proses tokenisasi dilakukan secara otomatis dengan menerapkan parameter *padding* dan *truncation* serta menetapkan *max length* 128, guna memastikan panjang input seragam dan efisien dalam pemrosesan.

3.4. Pelabelan Data

Tahap selanjutnya adalah pelabelan data untuk keperluan klasifikasi sentimen. Pada penelitian ini, label yang digunakan terdiri dari tiga kategori, yaitu positif, negatif, dan netral, yang merepresentasikan kecenderungan opini dalam setiap tweet. Tweet dengan sentimen positif menunjukkan dukungan, antusiasme, atau pujian terhadap topik yang dibahas (#BTScomeback), sedangkan tweet dengan sentimen negatif mengandung kritik, kekecewaan, atau respons emosional yang cenderung tidak mendukung. Sementara itu, tweet dengan sentimen netral merupakan opini yang bersifat informatif, deskriptif, atau tidak menunjukkan emosi secara eksplisit.

3.5. Split Data

Pada tahap ini, data yang digunakan sebanyak 6.300 tweet berbahasa Indonesia dengan tagar #BTScomeback, yang telah melalui proses pelabelan sentimen ke dalam tiga kategori, yaitu *positif*, *netral*, dan *negatif*. Sebelum dilakukan pembagian, seluruh data terlebih dahulu dibersihkan menggunakan teknik *preprocessing* yang mencakup penghapusan URL, mention, hashtag, angka, simbol, serta karakter non-ASCII. Selanjutnya,

data dibagi menjadi dua bagian menggunakan fungsi *train test split* dari pustaka *scikit-learn*, dengan dua skenario proporsi, yaitu 70:30 dan 80:20. Pada skenario 70:30, sebanyak 4.410 data digunakan sebagai data latih dan 1.890 data sebagai data uji. Sedangkan pada skenario 80:20, sebanyak 5.040 data digunakan sebagai data latih dan 1.260 data sebagai data uji. Untuk menjaga keseimbangan distribusi label sentimen pada kedua subset data, digunakan teknik *stratified sampling*. Selain itu, parameter *random state* 42 diterapkan untuk memastikan reprodusibilitas dan konsistensi hasil pembagian. Pembagian data yang terkontrol ini bertujuan agar model dapat belajar secara optimal pada data latih dan diuji secara adil pada data uji, sehingga mendukung proses evaluasi performa model secara objektif dan sistematis.

3.6. *Fine Tuning*

Pada tahap *fine-tuning*, model pra-latih IndoBERTweet diadaptasi khusus untuk tugas klasifikasi sentimen tiga kelas pada korpus tweet berbahasa Indonesia bertagar #BTScomeback. Model dasar dipasangi lapisan klasifikasi yang menghasilkan tiga skor (logit) sesuai jumlah kategori sentimen. Seluruh bobot model kemudian dioptimalkan ulang menggunakan data latih yang telah diproses menjadi representasi numerik (token ID dan *attention mask*) beserta label sentimennya. Proses pelatihan mengombinasikan *forward pass* untuk menghitung *loss*, *backward pass* untuk memperbarui bobot, dan evaluasi berkala pada data validasi setelah setiap epoch, dengan metrik utama akurasi. Pengaturan eksperimen meliputi penggunaan *batch size* yang sesuai, laju pembelajaran kecil (order 10^{-5}) dengan skema penurunan linier, dan regularisasi *weight decay* untuk mencegah *overfitting*. Selain itu, mekanisme *early stopping* diaktifkan sehingga pelatihan berhenti otomatis apabila tidak terjadi peningkatan akurasi validasi dalam dua epoch berturut-turut.

3.7. Evaluasi Model

Evaluasi Model dilakukan dengan kinerja model yang telah di *fine tuning* dievaluasi menggunakan data uji sebanyak 1.260 sampel yang sebelumnya belum pernah dilihat selama pelatihan. Setiap tweet diuji dengan memproses input melalui *tokenizer* dan model IndoBERTweet untuk menghasilkan prediksi kelas sentimen positif, netral, atau negatif yang kemudian dibandingkan dengan label asli. Pengukuran kinerja dilakukan dengan menghitung metrik *precision*, *recall*, dan *F1-score* untuk masing-masing kelas, serta akurasi keseluruhan, menggunakan fungsi *classification_report* dari pustaka scikit-learn.

4. HASIL DAN PEMBAHASAN

Bagian ini menyajikan rangkuman hasil penelitian yang telah dilakukan terkait analisis sentimen terhadap fenomena comeback BTS di platform X menggunakan model IndoBERTweet. Pada proses ini, data tweet berbahasa Indonesia yang terkumpul melalui teknik crawling dianalisis untuk mengukur pola opini dan emosional masyarakat Indonesia terhadap kembalinya grup musik tersebut.

4.1 Pra-Pemrosesan Data

Dalam proses analisis sentimen atau klasifikasi teks berbasis media sosial seperti Twitter, pra-pemrosesan data merupakan tahapan yang sangat penting untuk meningkatkan kualitas data sebelum dimasukkan ke dalam model analitik atau pembelajaran mesin. studi yang sedang dilakukan berfokus pada analisis sentimen terhadap *comeback* grup musik BTS, dengan memanfaatkan 6.300 data tweet yang telah dikumpulkan. Namun, sebelum data tersebut dapat dianalisis lebih lanjut, diperlukan tahapan pra-pemrosesan data yang sistematis yaitu *cleaning, lowercasing,* dan *stopword.*

4.1.1 *Cleaning*

Implementasi metode *cleaning* pada Gambar 3 menghasilkan teks yang telah dibersihkan dari karakter khusus dan simbol tidak diinginkan, namun tetap mempertahankan struktur kalimat yang dapat dibaca dengan baik. Dataset yang digunakan berisi informasi mengenai HYBE CEO dan persiapan comeback BTS 2025, yang diambil pada tanggal 19 April dengan conversation ID 1.91349E+18. Pada contoh data, hasil preprocessing menunjukkan teks "hybe ceo persiapan comeback butuh waktu kontrak dan visi grup jadi pertimbangan utama selengkapnya kgaeul" yang masih memiliki alur informasi yang koheren dan mudah dipahami.

created_at	conversation_id_str	full_text	cleaned_only
Sat Apr 19 06:54:37 +0000 2025	1.91349E+18	HYBE CEO: Persiapan Comeback #BTS 2025 Butuh Waktu Kontrak dan Visi Grup Jadi Pertimbangan Utama #HYBELabels #ComebackBTS #ARMY #Kpop Selengkapnya https://t.co/nZbXhpQ5By - K.Gaeul https://t.co/ifm73rCdSt	hybe ceo persiapan comeback butuh waktu kontrak visi grup jadi pertimbangan utama selengkapnya kgaeul

Gambar 3. *Cleaning Text*

4.1.2 *Lowercasing*

Penerapan metode *lowercased* pada Gambar 4 berhasil mengkonversi seluruh karakter alfabetik menjadi huruf kecil, menghasilkan normalisasi teks yang konsisten. Hasil preprocessing menampilkan format "hybe ceo persiapan comeback #bts 2025 butuh waktu kontrak dan visi grup jadi pertimbangan utama #hybelabels #comebackbts #army #kpop yang menunjukkan standardisasi penulisan tanpa mempengaruhi makna informasi. Normalisasi huruf kecil menghilangkan variasi penulisan yang disebabkan oleh perbedaan kapitalisasi, sehingga meningkatkan akurasi dalam proses matching dan indexing data. Metode ini sangat relevan untuk sistem pencarian informasi dan aplikasi yang memerlukan standardisasi format teks.

created_at	conversation_id_str	full_text	lowercased_only
Sat Apr 19 06:54:37 +0000 2025	1.91349E+18	HYBE CEO: Persiapan Comeback #BTS 2025 Butuh Waktu Kontrak dan Visi Grup Jadi Pertimbangan Utama #HYBELabels #ComebackBTS #ARMY #Kpop Selengkapnya https://t.co/nZbXhpQ5By - K.Gaeul https://t.co/ifm73rCdSt	hybe ceo: persiapan comeback #bts 2025 butuh waktu kontrak dan visi grup jadi pertimbangan utama #hybelabels #comebackbts #army #kpop selengkapnya https://t.co/nzbxhpq5by - k.gaeul https://t.co/ifm73rcdst

Gambar 4. Proses *Lowercasing*

4.1.3 *Stopword*

Proses *stopword* pada Gambar 5 menghasilkan ekstraksi kata kunci yang efektif dengan menghilangkan kata-kata penghubung yang tidak memberikan nilai informasi substantif. Output preprocessing menampilkan "hybe ceo persiapan *comeback* butuh waktu kontrak visi grup jadi pertimbangan utama selengkapnya kgaeul" yang lebih ringkas namun tetap mempertahankan esensi informasi utama.

created_at	conversation_id_str	full_text	stopwords_removed_only
Sat Apr 19 06:54:37 +0000 2025	1.91349E+18	HYBE CEO: Persiapan Comeback #BTS 2025 Butuh Waktu Kontrak dan Visi Grup Jadi Pertimbangan Utama #HYBELabels #ComebackBTS #ARMY #Kpop Selengkapnya https://t.co/nZbXhpQ5By - K.Gaeul https://t.co/ifm73rCdSt	hybe ceo persiapan comeback butuh waktu kontrak visi grup jadi pertimbangan utama selengkapnya kgaeul

Gambar 5. Proses *Stopword*

4.2. Tokenisasi

Tahap tokenisasi dan encoding dilakukan untuk mengubah teks mentah menjadi format numerik yang dapat dibaca oleh model. Tokenisasi dilakukan dengan memecah kalimat menjadi unit-unit kata atau simbol yang disebut token, menggunakan tokenizer khusus IndoBERTweet yang dirancang untuk memahami struktur bahasa Indonesia informal seperti slang dan singkatan. Setiap input teks diawali dengan token khusus [CLS] dan diakhiri dengan [SEP], sesuai standar arsitektur BERT. Selain itu, padding ([PAD]) diterapkan untuk menyamakan panjang input. Setelah proses tokenisasi, setiap token diubah menjadi angka melalui encoding berdasarkan kamus token (*vocabulary*) yang digunakan oleh model. Sebagai contoh, token “media” akan direpresentasikan dengan angka 3061, “mengatakan” dengan 2355, dan seterusnya. Hasil akhir dari proses ini adalah urutan bilangan bulat yang menjadi representasi dari teks untuk diproses lebih lanjut dalam tahap pelatihan dan inferensi oleh model IndoBERTweet seperti pada Gambar 6.

Input	Tokenized	Encoded
media mengatakan kembalinya bts bukan sekadar comeback seorang artis tapi merupakan reset untuk selu...	[CLS], media, mengatakan, kembalinya, bts, bukan, sekadar, comeback, seorang, artis, tapi, merupakan, rese, ##t, [SEP] ...	3, 3061, 2355, 15346, 29243, 2054, 6563, 28776, 1822, 5516, 2167, 1709, 25114, 930, 4 ...

Gambar 6. Tokenisasi dan *Encoding*

4.3. Pelabelan Data

Implementasi model klasifikasi terhadap data yang telah dipreprocessing menghasilkan prediksi label yang menunjukkan kategori sentimen atau topik dari konten teks. Salah satu sampel data pada Gambar 7 menunjukkan bahwa sistem berhasil mengklasifikasikan teks mengenai "HYBE CEO: Persiapan Comeback #BTS 2025 Butuh Waktu Kontrak dan Visi Grup Jadi Pertimbangan Utama #HYBELabels #ComebackBTS #ARMY #kpop

Selengkapnya" dengan label prediksi "netral". Hasil klasifikasi ini mencerminkan kemampuan model dalam mengidentifikasi tone dan karakteristik konten yang bersifat informatif tanpa bias emosional yang kuat. Konten yang membahas aspek bisnis dan perencanaan strategis dalam industri hiburan cenderung dikategorikan sebagai netral karena sifatnya yang faktual dan tidak mengandung sentimen positif atau negatif yang dominan.

created_at	conversation_id_str	full_text	predicted_label
Sat Apr 19 06:54:37 +0000 2025	1.91349E+18	HYBE CEO: Persiapan Comeback #BTS 2025 Butuh Waktu Kontrak dan Visi Grup Jadi Pertimbangan Utama #HYBELabels #ComebackBTS #ARMY #Kpop Selengkapnya https://t.co/nZbXhpQSBy - K.Gaeul https://t.co/ifm73rCdSt	netral

Gambar 7. Proses Pelabelan Data

4.4. Visualisasi word cloud

Visualisasi *word cloud* digunakan sebagai metode eksploratif guna menampilkan frekuensi kemunculan kata dalam kumpulan tweet bertagar #BTScomeback. Setiap kata ditampilkan dengan ukuran yang proporsional terhadap jumlah kemunculannya dalam data, sehingga semakin sering sebuah kata muncul, maka semakin besar pula tampilannya pada *word cloud.*

Gambar 8. *Word cloud* sentimen positif

Pada Gambar 8. ditampilkan representasi visual dari kata-kata yang paling sering muncul dalam tweet berlabel sentimen positif. *Word cloud* ini menunjukkan bahwa ekspresi publik terhadap tagar #BTScomeback didominasi oleh kata-kata yang mencerminkan antusiasme, dukungan, dan harapan yang kuat terhadap kembalinya BTS ke panggung hiburan. Beberapa kata seperti "*comeback*", "bts", "lagi", "banget", "buat", dan "aku" muncul dengan ukuran yang menonjol, menandakan frekuensi kemunculan yang tinggi dalam tweet positif. Kehadiran kata-kata tersebut mencerminkan respons emosional positif yang kuat dari para pengguna terhadap momen comeback BTS. Selain itu, ungkapan seperti "semoga", "sukses", "semangat", dan "yukkk" juga muncul dan memperkuat konteks dukungan serta semangat kolektif dari para penggemar. Kata ganti orang seperti "aku", "kita", dan "mereka" turut memperlihatkan adanya nuansa kebersamaan dan interaksi sosial, yang mempertegas bahwa antusiasme terhadap BTS tidak hanya bersifat individual, tetapi juga menjadi pengalaman bersama dalam komunitas.

Gambar 9. *World cloud* sentimen negatif

Gambar 9. menyajikan visualisasi word cloud dari tweet yang tergolong dalam kategori sentimen negatif terhadap tagar #BTScomeback. *Word cloud* ini memperlihatkan sejumlah kata yang sering muncul dalam tweet bernada negatif, yang secara umum mencerminkan ekspresi kekecewaan, sindiran, hingga kritik terhadap momen comeback BTS. Beberapa kata yang dominan seperti “comeback”, “itu”, “nggak”, “gw”, “udah”, “lu”, dan “gak” menunjukkan adanya respons skeptis atau kurang antusias dari sebagian pengguna. Selain itu, kehadiran kata-kata seperti “cape”, “sendiri”, “takut”, “kenapa”, dan “sibuk” mengindikasikan adanya perasaan jenuh, kecewa, atau konflik emosional tertentu terhadap fenomena comeback tersebut. Penggunaan kata ganti orang pertama dan kedua seperti “aku”, “gw”, “lu”, serta bentuk informal lainnya menunjukkan nuansa personal dalam penyampaian opini negatif. Hal ini menandakan bahwa ekspresi ketidakpuasan kerap disampaikan secara langsung dan spontan, sesuai karakter platform media sosial seperti X (Twitter). Secara keseluruhan, *word cloud* ini merepresentasikan dinamika opini publik yang bersifat kritis atau negatif terhadap

#BTScomeback. Visualisasi ini memperkuat pemahaman bahwa tidak semua reaksi terhadap fenomena budaya populer bersifat positif, dan analisis sentimen mampu menangkap spektrum emosi tersebut secara lebih luas dan kontekstual.

Gambar 10. *Word Cloud* sentimen netral

Gambar 10 menunjukkan distribusi kata-kata yang sering muncul dalam tweet dengan sentimen netral terkait tagar #BTScomeback. Kata-kata seperti "comeback", "bts", "ini", "udah", dan "lagi" tampak mendominasi, mencerminkan gaya bahasa informatif tanpa muatan emosi yang kuat. Kata-kata ini cenderung menyampaikan fakta atau opini netral seputar jadwal, aktivitas, atau informasi umum mengenai BTS. Beberapa kata seperti "mau", "bisa", "lihat", dan "lagu" menunjukkan adanya ekspresi harapan atau perhatian tanpa penekanan emosional tertentu. Kata "*army*", "kita", dan "kak" juga muncul sebagai bagian dari interaksi komunitas, namun dalam konteks yang tidak terlalu emosional. *Word cloud* ini mengindikasikan bahwa tidak semua respons publik bersifat emosional; sebagian hanya berisi penyampaian informasi atau opini secara netral. Visualisasi ini membantu menggambarkan isi tweet yang bersifat deskriptif atau naratif tanpa kecenderungan ke arah sentimen positif maupun negatif. Hal ini penting untuk menunjukkan keberagaman ekspresi publik yang tidak selalu dipengaruhi oleh emosi.

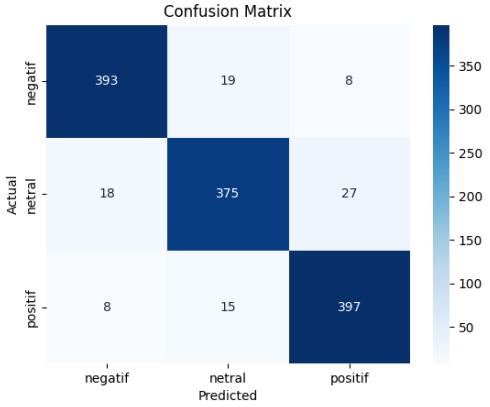
Gambar 11. *Word Cloud* semua kata

Word cloud pada Gambar 11 memberikan gambaran menyeluruh mengenai frekuensi kemunculan kata-kata dalam seluruh kumpulan tweet yang dianalisis, tanpa dibedakan berdasarkan kategori sentimen. Kata-kata seperti "bts", "comeback", "aku", "lagi", "ini", dan "dan" menjadi yang paling menonjol, mencerminkan intensitas pembahasan publik terhadap topik utama yaitu comeback grup BTS. Dominasi kata "bts" dan "comeback" menegaskan bahwa percakapan yang terjadi memang sangat terpusat pada isu tersebut. Kemunculan kata "army", "mereka", "buat", dan "bangtan" turut memperlihatkan konteks percakapan yang erat kaitannya dengan fandom BTS, serta menunjukkan dimensi sosial dari komunikasi yang terjadi di platform X. Sementara itu, kata-kata seperti "tapi", "udah", "kita", "mau", dan "jadi" memberi indikasi adanya keragaman opini dan ekspresi pengguna yang bisa bernada positif, negatif, maupun netral. Beberapa kata netral seperti "yang", "itu", "ya", "dulu", dan "gitu" juga muncul dengan frekuensi tinggi karena sifatnya yang umum dalam struktur kalimat bahasa Indonesia. Hal ini mencerminkan karakteristik linguistik khas pengguna lokal yang menggunakan bahasa informal dalam menyampaikan pendapat. Secara keseluruhan, *word cloud* ini menunjukkan bahwa pembahasan seputar #BTScomeback melibatkan berbagai nuansa emosi dan ekspresi, dengan intensitas tinggi pada kata-kata yang mencerminkan antusiasme, ekspektasi, serta keterlibatan komunitas penggemar dalam merespons fenomena comeback BTS di awal hingga pertengahan tahun 2025. Visualisasi ini menjadi fondasi penting dalam memahami pola-pola linguistik dan fokus utama dari percakapan digital yang terjadi selama periode pengumpulan data.

4.5. Evaluasi Model

Evaluasi performa dilakukan terhadap dua model klasifikasi sentimen berbasis IndoBERTweet yang telah di-fine-tune menggunakan data uji berimbang dari tiga kategori sentimen, yaitu negatif, netral, dan positif. Model pertama dengan perbandingan 80:20 diuji menggunakan 1.260 data, sedangkan model kedua dengan perbandingan 70:30 menggunakan 1.890 data. Evaluasi ini dilakukan sebagai tolak

ukur sejauh mana model mampu mengenali dan membedakan ketiga kategori sentimen secara akurat. Pada gambar 12 dan 13 diperoleh akurasi keseluruhan sebesar 92%. Nilai precision, recall, dan f1-score tertinggi diperoleh pada kelas negatif dan positif, yakni masing-masing mencapai 0.94 dan 0.93. Performa model pada kelas netral relatif lebih rendah, dengan nilai f1-score sebesar 0.89. Hal ini menunjukkan bahwa model masih mengalami kesulitan dalam membedakan tweet yang bersifat netral, yang sering kali bersifat ambigu atau kontekstual. *Confusion matrix* juga menunjukkan adanya misclassification yang cukup besar pada kelas ini, yaitu sebanyak 51 tweet salah diklasifikasikan ke kelas negatif maupun positif.



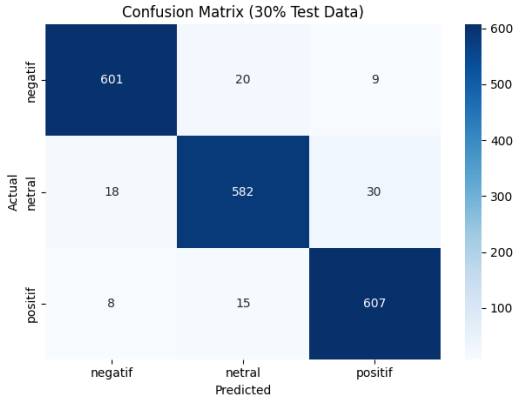
Gambar 12. *Confusion matrix* 80:20

Classification Report:

	precision	recall	f1-score	support
negatif	0.94	0.94	0.94	420
netral	0.92	0.89	0.90	420
positif	0.92	0.95	0.93	420
accuracy			0.92	1260
macro avg	0.92	0.92	0.92	1260
weighted avg	0.92	0.92	0.92	1260

Gambar 13. *Clasification Report* 80:20

Gambar 14 dan 15 menunjukkan performa yang lebih unggul secara keseluruhan, dengan akurasi mencapai 95%. F1-score tertinggi berada pada kelas negatif sebesar 0.96 dan positif sebesar 0.95, serta pada kelas netral sebesar 0.93. Nilai-nilai ini menunjukkan bahwa model tidak hanya akurat dalam prediksi, tetapi juga seimbang dalam hal sensitivitas dan presisi antar kelas. *Confusion matrix* menunjukkan bahwa sebagian besar tweet berhasil diklasifikasikan dengan tepat, dengan jumlah kesalahan klasifikasi yang jauh lebih kecil dibandingkan model sebelumnya. Peningkatan performa ini mencerminkan kemampuan model dalam mengenali variasi ekspresi sentimen yang lebih kompleks, termasuk dalam mengidentifikasi opini yang bersifat netral.



Gambar 14. *Confusion Matrix* 70:30

Classification Report:

	precision	recall	f1-score	support
negatif	0.96	0.95	0.96	630
netral	0.94	0.92	0.93	630
positif	0.94	0.96	0.95	630
accuracy			0.95	1890
macro avg	0.95	0.95	0.95	1890
weighted avg	0.95	0.95	0.95	1890

Gambar 15. *Classification Report* 70:30

Tabel 4 menunjukkan bahwa pembagian data dengan rasio 70:30 menghasilkan performa model yang lebih baik dibandingkan dengan rasio 80:20. Model dengan rasio 70:30 mampu mencapai akurasi sebesar 95%, atau lebih tinggi 3% dibandingkan model dengan rasio 80:20 yang hanya mencapai 92%. Peningkatan ini mengindikasikan bahwa alokasi data uji yang lebih besar memungkinkan evaluasi model yang lebih representatif dan menyeluruh. Dengan proporsi data uji yang lebih besar, model tidak hanya mengalami proses evaluasi

yang lebih ketat, tetapi juga menunjukkan kemampuan yang lebih unggul dalam mengenali variasi ekspresi sentimen secara menyeluruh, termasuk sentimen netral yang dikenal lebih kompleks dan ambigu. Temuan ini menegaskan bahwa pemilihan strategi pembagian data memiliki peran penting dalam mengoptimalkan performa model IndoBERTweet, khususnya dalam konteks klasifikasi sentimen terhadap teks berbahasa Indonesia di media sosial.

Tabel 4. Perbandingan hasil split data

Pembagian Data	Akurasi
70:30	**95%**
80:20	92%

5. KESIMPULAN

Hasil penelitian menunjukkan bahwa model IndoBERTweet mampu mengklasifikasikan sentimen publik terhadap fenomena comeback BTS di platform X ke dalam tiga kelas utama, yaitu positif, netral, dan negatif, dengan performa yang sangat baik. Dengan total 6.300 data yang dianalisis, data dibagi menggunakan rasio 70:30, yakni 70% (4.410 data) sebagai data latih dan 30% (1.890 data) sebagai data uji. Model mencatat akurasi tertinggi sebesar 95%, menandakan efektivitas model dalam menangani bahasa informal khas media sosial dan mengidentifikasi ekspresi sentimen secara akurat di seluruh kelas. Evaluasi performa memperlihatkan kinerja model yang optimal, terutama dalam mendeteksi sentimen positif dan negatif, serta menunjukkan peningkatan yang signifikan pada kelas netral yang cenderung lebih kompleks. Keberhasilan ini juga didukung oleh alokasi data uji yang lebih besar, yang memungkinkan proses evaluasi yang lebih representatif terhadap kemampuan generalisasi model. Temuan ini menegaskan bahwa IndoBERTweet merupakan pendekatan yang tepat untuk analisis sentimen berbasis teks media sosial berbahasa Indonesia dalam konteks budaya populer. Untuk pengembangan selanjutnya, disarankan perluasan cakupan data dan integrasi teknik analisis emosi yang lebih mendalam guna memperkuat akurasi dan kedalaman pemahaman terhadap opini publik digital di Indonesia.

DAFTAR PUSTAKA

[1] A. Puspita Sari, A. A. Widoretno, F. Prima Aditiawan, and A. M. Rizki, "Peningkatan Ekonomi Digital Pada Usaha Kerajinan Kulit Melalui Optimalisasi Teknologi Informasi," *Jurnal Pengabdian kepada Masyarakat Nusantara (JPkMN)*, vol. 6, no. 1.1, pp. 397–402, 2024.

[2] S. Suhendra and F. Selly Pratiwi, "Peran Komunikasi Digital Dalam Pembentukan Opini Publik: Studi Kasus Media Sosial," *Iapa Proceedings Conference*, p. 293, Oct. 2024, doi: 10.30589/proceedings.2024.1059.

[3] E. Putri Wardani, R. Sari Kusuma, U. Muhammadiyah Surakarta, J. Ahmad Yani, and J. Tengah, "Interaksi Parasosial Penggemar K-Pop Di Media Sosial (Studi Kualitatif Pada Fandom Army Di Twitter)," *Jurnal Magister Ilmu Komunikasi*, vol. 7, no. 2, pp. 243–260, [Online]. Available: http://journal.ubm.ac.id/

[4] M. Arsyad, "Analisis Sentimen Komputasional: Memahami Emosi Dalam Teks." [Online]. Available: https://www.researchgate.net/publication/376722063

[5] S. Mughal, A. Jaffar, and W. Arif, "Sentiment Analysis Of Social Media Data: Understanding Public Perception", doi: 10.56979/702/2024.

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

[7] F. Koto Jey Han Lau Timothy Baldwin, "IndoBERTweet: A Pretrained Languange Model For Indonesian Twitter With Effective Domain-Specific Vocabulary Initialization." [Online]. Available: https://huggingface.co/huseinzol05/

[8] R. A. Fitrianto and A. S. Editya, "Klasifikasi Tweet Sarkasme Pada Platform X Menggunakan Bidirectional Encoder Representations From Transformers," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 3, pp. 366–371, Jul. 2024, doi: 10.47233/jteksis.v6i3.1344.

[9] L. Widya Astuti and Y. Sari, "Code-Mixed Sentiment Analysis Using Transformer For Twitter Social Media Data." [Online]. Available: www.ijacsa.thesai.org

[10] S. Muspani, I. Made, A. D. Suarjaya, N. Kadek, and D. Rusjayanthi, "Sentiment Analysis Related To Korean Subculture In Indonesia Using BERT Method," 2024. [Online]. Available: https://t.co/U1HmmS7dDA

[11] M. T. Uliniansyah, A. Jarin, A. Santosa, and G. Gunarso, "Modeling Sentiment Analysis Of Indonesian Biodiversity Policy Tweets Using IndoBERTweet," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 3, p. 2389, Jun. 2025, doi: 10.11591/ijai.v14.i3.pp2389-2401.

[12] E. Fachrosi *et al.*, "Dinamika Fanatisme Penggemar K-Pop Pada Komunitas BTS-Army Medan The Dynamics of K-Pop Fanaticism In The BTS-Army Medan Community," *Jurnal Diversita*, vol. 6, no. 2, 2020, doi: 10.31289/diversita.v6i2.3782.

[13] D. Gloria Masada, "Cultural Fusion: The Impact Of K-Pop On The Indonesian Perspectives And Identities", doi: 10.26593/sentris.v5i1.7633.34-44.

[14] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," vol. 8, no. 1, pp. 147–156, 2021, doi: 10.25126/jtiik.202183944.

[15] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, Jan. 2022, doi: 10.23960/jitet.v10i1.2262.

[16] T. Safitri, Y. Umaidah, and I. Maulana, "Analisis Sentimen Pengguna Twitter Terhadap BTS Menggunakan Algoritma Support Vector Machine," 2023. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[17] Z. A. Sriyanti, D. S. Y. Kartika, and A. R. E. Najaf, "Implementasi Model Bert Pada Analisis Sentimen Pengguna Twitter Terhadap Aksi Boikot Produk Israel," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4743.

[18] Y. A. Singgalen, "Performance Analysis Of IndoBERT For Sentiment Classification In Indonesian Hotel Review Data," *Article in Journal of Information System Research*, vol. 6, no. 2, 2025, doi: 10.47065/josh.v6i2.6505.

[19] H. M. Lee and Y. Sibaroni, "Comparison Of IndoBERTweet and Support Vector Machine On Sentiment Analysis Of Racing Circuit Construction In Indonesia," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 1, p. 99, Jan. 2023, doi: 10.30865/mib.v7i1.5380.

[20] S. Sato, J. Sakuma, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa, "Vocabulary Adaptation For Domain Adaptation in Neural Machine Translation."

[21] M. Fadhel, "Depression Detection Of Users In Social Media X Using IndoBERTweet," *Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, 2024, doi: 10.33395/v8i2.13354.