

Comparative Analysis of IndoRoBERTa and EmoRoBERTa for Emotion Classification Based on Ekman's Taxonomy Using the Indonesian GoEmotions Dataset

Syifa Naura Milla Celesta¹, Dinda Lestari^{2*}

* Sistem Informasi, Universitas Sriwijaya

09031282328025@student.unsri.ac.id¹, dinda@unsri.ac.id²

Article Info

Article history:

Received 2026-09-08

Revised 2026-10-14

Accepted 2026-10-18

Keyword:

*Emotion Classification,
EmoRoBERTa,
GoEmotions,
IndoRoBERTa,
Ekman's Taxonomy*

ABSTRACT

Emotion classification is an important task in Natural Language Processing (NLP) for identifying emotional information conveyed through Indonesian-language text. However, differences in model specialization may lead to different classification performance, while direct comparisons between language-oriented and emotion-oriented models under the same experimental setting remain limited. This study compares the performance of IndoRoBERTa and EmoRoBERTa for emotion classification using the Indonesian version of the GoEmotions dataset, with mBERT and TF-IDF with SVM included as baseline approaches. The dataset consists of 45,691 single-label samples categorized into seven emotions based on Ekman's Taxonomy, namely *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *neutral*. The experimental process includes text preprocessing, class weighting to address class imbalance, and hyperparameter tuning to determine the optimal configuration for each model. Model performance is evaluated using *accuracy*, *macro precision*, *macro recall*, and *macro F1-score*, with *macro F1-score* as the primary metric. Under the dataset and experimental settings used in this study, IndoRoBERTa obtains an *accuracy* of 0.6603, *macro precision* of 0.5938, *macro recall* of 0.5975, and *macro F1-score* of 0.5887, while EmoRoBERTa obtains 0.6326, 0.5613, 0.5956, and 0.5742, respectively. mBERT and TF-IDF with SVM obtain *macro F1-scores* of 0.5495 and 0.5449, respectively. The results provide an empirical comparison of language-oriented and emotion-oriented Transformer models within the experimental setting and provide baseline context for interpreting their classification performance.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. PENDAHULUAN

Seiring dengan pesatnya perkembangan media sosial dan berbagai platform digital, masyarakat semakin aktif mengekspresikan pendapat dan pengalamannya melalui teks dalam bentuk komentar. Komentar-komentar tersebut tidak hanya merepresentasikan keragaman opini dan tanggapan terhadap berbagai isu, tetapi juga mencerminkan emosi yang mendasari respons pengguna terhadap suatu peristiwa atau topik tertentu [1]. Melimpahnya data teks yang mengandung ekspresi emosi menjadikan analisis emosi semakin penting karena mampu memperluas pemahaman mengenai cara pengguna merespons suatu informasi atau peristiwa. Oleh

sebab itu, kemampuan mengenali emosi memiliki peran yang cukup besar, baik untuk mendukung interaksi antarmanusia maupun kerja sama antara manusia dan mesin. Dalam perkembangan penelitian terkini, *Natural Language Processing* (NLP) yang didukung oleh kemajuan *deep learning* dan *powerful language models* telah mendorong pesatnya penelitian di bidang *emotion recognition* [2]. Kemampuan mengenali emosi tersebut telah dimanfaatkan pada berbagai aplikasi, seperti analisis media sosial, *conversational AI*, *affective computing*, dan pemantauan kesehatan mental. Namun, emosi dalam teks masih menjadi tantangan untuk dikenali karena sangat bergantung pada

konteks, sering kali bersifat implisit, serta dipengaruhi oleh faktor budaya [2].

Dalam upaya mengenali emosi pada data teks, pendekatan yang paling awal dan banyak digunakan adalah *sentiment analysis*, yaitu metode yang mengelompokkan teks berdasarkan polaritas positif, negatif, atau netral [3]. Meskipun efektif untuk menggambarkan kecenderungan sentimen suatu teks, pendekatan tersebut belum mampu merepresentasikan keragaman dan kompleksitas emosi manusia secara menyeluruh karena hanya mengelompokkan teks berdasarkan polaritas umum [3]. Akibatnya, pendekatan tersebut masih memiliki keterbatasan dalam mengenali berbagai kategori emosi, termasuk senang, sedih, gembira, takut, serta cinta. Keterbatasan tersebut mendorong perkembangan penelitian menuju klasifikasi emosi yang berfokus pada identifikasi emosi secara lebih spesifik dalam sebuah teks. Proses klasifikasi emosi bertujuan untuk menentukan kategori emosi yang terdapat pada teks, baik berupa kalimat maupun paragraf [4]. Melalui proses tersebut, sistem mampu memprediksi emosi dari data teks secara lebih akurat sehingga memberikan pemahaman yang lebih mendalam mengenai ekspresi emosional yang terkandung dalam teks serta mendukung pengambilan keputusan yang lebih tepat.

Perkembangan penelitian di bidang klasifikasi emosi turut mendorong pengembangan berbagai dataset yang dirancang untuk merepresentasikan emosi manusia secara lebih komprehensif. Dalam penelitian klasifikasi emosi, GoEmotions banyak digunakan sebagai dataset *benchmark*. Dataset ini terdiri dari 58.000 komentar pengguna dari platform media sosial reddit yang dianotasi ke dalam 27 kategori emosi dan satu kategori netral [5]. Selain menyediakan anotasi emosi secara *fine-grained*, GoEmotions juga menyediakan pemetaan kategori emosi ke dalam beberapa taksonomi, termasuk taksonomi Ekman [5]. Pada penelitian ini digunakan dataset GoEmotions yang telah tersedia dalam terjemahan bahasa Indonesia dengan pendekatan *single-label* berdasarkan pemetaan taksonomi Ekman sehingga menghasilkan tujuh kategori emosi. Karakteristik data tersebut memerlukan penggunaan model yang dapat menangkap konteks kebahasaan sekaligus mengidentifikasi karakteristik ekspresi emosi di dalam teks. Oleh karena itu, berbagai *pre-trained language* model berbasis transformer mulai banyak dimanfaatkan dalam penelitian klasifikasi emosi karena kemampuannya dalam memahami representasi bahasa secara kontekstual.

Model berbasis transformer telah menunjukkan kemampuan yang baik dalam menyelesaikan tugas klasifikasi emosi pada berbagai penelitian terdahulu. Pada pengembangan awal dataset GoEmotions, Demszky dkk. [5] mengevaluasi model berbasis BERT dan memperoleh rata-rata F1-score sebesar 0.46. Capaian tersebut memperlihatkan bahwa klasifikasi emosi masih menyisakan ruang yang luas untuk peningkatan performa. Seiring dengan perkembangan penelitian, berbagai arsitektur transformer kemudian dikembangkan untuk mengoptimalkan kemampuan model

dalam memahami representasi emosional di dalam teks. Penelitian yang dilakukan oleh Kamath dkk. [6] dengan mengembangkan model EmoRoBERTa, yaitu model RoBERTa yang di-fine-tuning menggunakan dataset GoEmotions sehingga secara khusus dioptimalkan untuk memahami representasi emosi dalam teks. Model tersebut dievaluasi pada tiga tingkat taksonomi emosi, yaitu sentiment groups, taksonomi Ekman, dan taksonomi GoEmotions, serta menghasilkan macro F1-score yang lebih tinggi dibandingkan model pada penelitian awal GoEmotion, yaitu sebesar 0.49.

Model dasar RoBERTa dikembangkan melalui proses *pre-training* menggunakan korpus berbahasa Inggris, sedangkan GoEmotions yang digunakan dalam proses *fine-tuning* EmoRoBERTa juga terdiri atas teks berbahasa Inggris [7] [8]. Dalam merepresentasikan input, RoBERTa menggunakan *byte-level Byte Pair Encoding* (BPE) yang merepresentasikan teks melalui unit-unit *subword* [9]. Karakteristik tersebut memungkinkan teks direpresentasikan melalui unit-unit *subword* yang membentuk kata, sehingga dapat membantu menangani kata yang tidak tersedia dalam *vocabulary* sebagai satu kesatuan [9] [10]. Penggunaan arsitektur RoBERTa dengan *byte-level BPE* juga telah diterapkan pada pemrosesan teks berbahasa Indonesia dalam penelitian terdahulu, menggunakan data berbahasa Indonesia untuk pengembangan sistem *question answering* [11]. Dengan demikian, ketika diterapkan pada teks berbahasa Indonesia, EmoRoBERTa membawa spesialisasi pada representasi emosi yang diperoleh melalui proses *fine-tuning*, sementara representasi bahasa dasarnya berasal dari model RoBERTa yang dikembangkan menggunakan data berbahasa Inggris. Karakteristik tersebut menunjukkan bahwa penerapan EmoRoBERTa pada teks berbahasa Indonesia memiliki perbedaan dengan model yang sejak awal dikembangkan menggunakan data berbahasa Indonesia.

Perbedaan karakteristik tersebut dapat dilihat pada penelitian klasifikasi emosi menggunakan model yang dikembangkan untuk bahasa Indonesia. Pada konteks bahasa Indonesia, Putri dkk. [12] membandingkan performa IndoBERT dan DistilBERT menggunakan dataset GoEmotions versi bahasa Indonesia, dengan hasil IndoBERT memperoleh nilai *macro* F1-score 0.4659, sedikit lebih tinggi dibandingkan DistilBERT yang hanya mencapai nilai 0.4414. Selain itu, penelitian oleh Lazuardi dkk. [13] menerapkan IndoRoBERTa pada tugas klasifikasi emosi multikelas menggunakan teks berbahasa Indonesia. Hasil pada penelitian tersebut menunjukkan bahwa IndoRoBERTa mampu mengidentifikasi kategori emosi pada teks berbahasa Indonesia. Temuan tersebut memperlihatkan bahwa klasifikasi emosi pada teks berbahasa Indonesia dapat dilakukan menggunakan model dengan orientasi spesialisasi yang berbeda, yaitu model yang dioptimalkan untuk merepresentasikan informasi emosional dan model yang dikembangkan untuk menangkap karakteristik bahasa Indonesia.

Selain model berbasis transformer, klasifikasi emosi juga telah dilakukan menggunakan pendekatan *machine learning*

konvensional. Salah satu pendekatan tersebut adalah kombinasi *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk representasi fitur teks dan *Support Vector Machine* (SVM) sebagai algoritma klasifikasi [14]. Penelitian Fitriana dan Setiawan [14] menggunakan kombinasi SVM dengan TF-IDF pada subset GoEmotions dan memperoleh akurasi sebesar 35%. Sementara itu, Zamsuri dkk. [15] menerapkan SVM dengan representasi TF-IDF pada teks berbahasa Indonesia dan memperoleh akurasi sebesar 62% setelah pengembangan konfigurasi kernel. Hasil tersebut menunjukkan bahwa pendekatan berbasis konvensional masih mampu digunakan untuk klasifikasi emosi, meskipun kinerjanya dapat berbeda bergantung pada karakteristik data dan konfigurasi model. Hasil ini memberikan titik awal untuk melihat bagaimana pendekatan berbasis representasi bahasa *pre-trained* dapat memberikan kinerja pada tugas yang sama.

Perkembangan *pre-trained language* model kemudian memungkinkan klasifikasi dilakukan dengan memanfaatkan representasi bahasa yang telah dipelajari sebelumnya. Salah satu pendekatan tersebut adalah Multilingual BERT (mBERT) yang dipra-latih menggunakan korpus Wikipedia dalam 104 bahasa, termasuk bahasa Indonesia [16]. Pada penelitian klasifikasi emosi berbahasa Indonesia, mBERT memperoleh nilai macro-F1 sebesar 0.633, sedangkan IndoRoBERTa memperoleh 0.694 pada dataset yang sama [17]. Hasil tersebut memberikan gambaran bahwa model multilingual dapat digunakan untuk klasifikasi emosi berbahasa Indonesia, tetapi model yang secara khusus dikembangkan untuk bahasa Indonesia dapat menunjukkan kinerja yang berbeda pada tugas dan data tertentu. Perbedaan tersebut menjadi relevan untuk melihat bagaimana karakteristik spesialisasi model berkaitan dengan kinerja klasifikasi emosi.

Perbedaan karakteristik tersebut menjadi semakin menarik ketika dibandingkan dengan model yang memiliki orientasi spesialisasi berbeda. IndoRoBERTa dikembangkan dengan fokus pada karakteristik bahasa Indonesia, sedangkan EmoRoBERTa dioptimalkan untuk merepresentasikan informasi emosional melalui proses *fine-tuning* pada GoEmotions. Kondisi tersebut menimbulkan pertanyaan mengenai orientasi spesialisasi yang lebih sesuai untuk klasifikasi emosi pada teks berbahasa Indonesia, yaitu apakah kemampuan yang lebih terarah pada karakteristik bahasa atau kemampuan yang lebih terarah pada representasi emosi. Perbandingan kedua orientasi tersebut menjadi penting untuk melihat bagaimana karakteristik spesialisasi model berkaitan dengan kemampuan model dalam mengklasifikasikan emosi pada teks berbahasa Indonesia.

Namun, kajian terdahulu umumnya masih menempatkan pengembangan maupun evaluasi masing-masing model secara terpisah sebagai fokus utama. Penelitian yang mengevaluasi EmoRoBERTa pada teks berbahasa Indonesia masih sangat terbatas. Di sisi lain, kajian yang membandingkan secara langsung model dengan spesialisasi pada karakteristik bahasa Indonesia dan model dengan spesialisasi pada representasi emosi menggunakan dataset

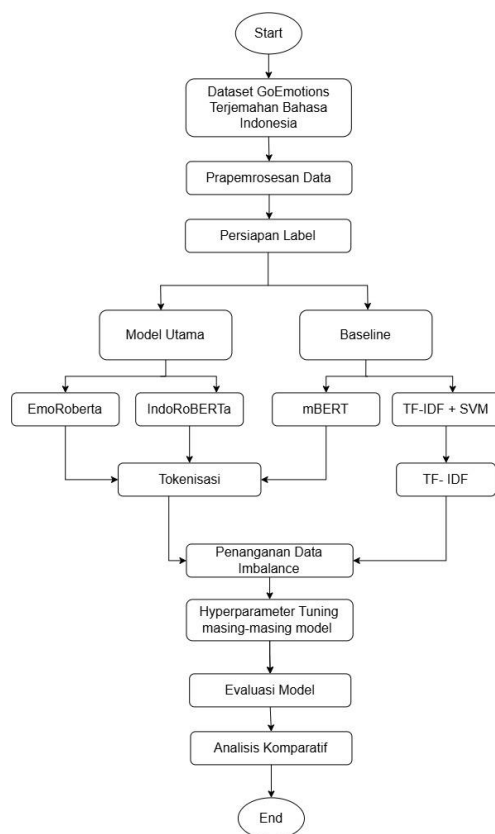
dan skema klasifikasi yang sama juga masih terbatas. Keterbatasan tersebut menyebabkan belum diketahui bagaimana perbedaan orientasi spesialisasi model, yaitu spesialisasi bahasa dan spesialisasi emosi, berkaitan dengan kinerja klasifikasi emosi pada teks berbahasa Indonesia ketika kedua model dievaluasi dalam kondisi eksperimen yang sama.

Berdasarkan kesenjangan tersebut, penelitian ini diarahkan untuk menganalisis kinerja model dengan orientasi spesialisasi yang berbeda dalam klasifikasi emosi pada teks berbahasa Indonesia, dengan memfokuskan perbandingan pada EmoRoBERTa sebagai model berorientasi representasi emosi dan IndoRoBERTa sebagai model yang dikembangkan untuk karakteristik bahasa Indonesia. Pendekatan SVM dengan TF-IDF serta model multilingual mBERT turut digunakan sebagai pembanding untuk memberikan konteks terhadap kinerja kedua model tersebut. Setiap model dioptimalkan melalui proses *hyperparameter tuning* sebelum dilakukan evaluasi pada data pengujian. Hasil penelitian diharapkan dapat memberikan gambaran empiris mengenai keterkaitan antara orientasi spesialisasi model dan kinerja klasifikasi emosi pada teks berbahasa Indonesia.

II. METODE

Metode eksperimen kuantitatif komparatif merupakan metode yang diterapkan pada penelitian ini, dengan membandingkan kinerja dua *pre-trained* model berbasis arsitektur RoBERTa, yaitu IndoRoBERTa dan EmoRoBERTa dalam melakukan klasifikasi emosi pada teks berbahasa Indonesia. Selain kedua model utama tersebut, mBERT dan SVM dengan TF-IDF digunakan sebagai model *baseline* untuk memberikan konteks pembanding. Evaluasi terhadap model-model tersebut dilakukan menggunakan dataset GoEmotions versi bahasa Indonesia yang disusun dalam skema *single-label* berdasarkan taksonomi Ekman.

Proses penelitian diawali dengan data *preprocessing* dan penanganan data *imbalance*, yang kemudian dilanjutkan dengan penerapan model *pre-trained*. Selanjutnya, *hyperparameter tuning* dilakukan untuk menentukan konfigurasi optimal dari masing-masing model. Evaluasi terhadap model-model tersebut kemudian diukur melalui *accuracy*, *macro precision*, *macro recall*, serta *macro F1-score*. Analisis terhadap hasil prediksi dilakukan dengan meninjau pola kesalahan pada setiap kategori emosi melalui *confusion matrix*, sedangkan *training loss* dan *validation loss* dianalisis pada kedua model utama guna mengamati dinamika pembelajaran selama tahap pelatihan. Hasil pengukuran tersebut selanjutnya digunakan sebagai dasar untuk membandingkan capaian kedua model utama, serta mBERT dan SVM dengan TF-IDF sebagai konteks pembanding dalam tahap pembahasan.



Gambar 1. Alur Penelitian

A. Dataset

Goemotions merupakan dataset yang digunakan dalam penelitian ini, yang diperkenalkan oleh Demszy dkk. [5] melalui penelitiannya sebagai dataset utama. Dataset tersebut terdiri atas sekitar 58.000 komentar berbahasa Inggris dari platform Reddit yang telah dianotasi ke dalam 27 kategori emosi dan satu kategori netral [5]. Dataset GoEmotions juga telah terbagi ke dalam tiga subset, yaitu pelatihan, validasi, dan pengujian, yang masing-masing mencakup 80%, 10%, dan 10% dari keseluruhan data. Pada penelitian ini digunakan versi terjemahan bahasa Indonesia dari dataset GoEmotions yang diperoleh melalui repositori GitHub dari penelitian sebelumnya [18]. Proses penerjemahan dilakukan secara otomatis dari bahasa Inggris ke bahasa Indonesia, meliputi teks komentar dan label emosi.

GoEmotions merupakan dataset multi-label, yaitu sebuah komentar dapat memuat beberapa label emosi. Namun, sebagian besar sampel pada dataset tersebut hanya memiliki satu label emosi [5]. Berdasarkan karakteristik tersebut, penelitian ini difokuskan pada data dengan satu label emosi agar proses klasifikasi lebih terfokus pada satu kategori emosi yang merepresentasikan setiap komentar. Setelah proses penyaringan, diperoleh sebanyak 45.691 data yang terdiri atas 36.572 data pelatihan, 4.538 data validasi, dan 4.581 data pengujian.

B. Prapemrosesan Data

1) Case Folding

Pada proses *case folding*, bentuk penulisan teks diseragamkan dengan merepresentasikan seluruh karakter dalam huruf kecil. Tahapan ini diterapkan agar variasi penggunaan huruf kapital tidak menghasilkan perbedaan representasi kata [19]. Penyeragaman tersebut memungkinkan kata yang memiliki makna sama memperoleh representasi token yang konsisten.

2) Text Cleaning

Dalam tahap *text cleaning*, kualitas data ditingkatkan dengan menghapus berbagai elemen teks yang tidak relevan [19]. Proses diawali dengan penghapusan data yang memiliki *missing value* dan data duplikat. Data duplikat dapat menyebabkan bias pada proses analisis maupun pelatihan model karena informasi yang sama muncul lebih dari satu kali, sehingga distribusi data menjadi kurang representatif serta meningkatkan risiko data *leakage* pada proses pelatihan dan evaluasi model [20]. Selanjutnya dilakukan penghapusan URL, *mention*, pembersihan *hashtag* serta karakter atau simbol yang tidak diperlukan.

3) Text Normalization

Text normalization dilakukan dengan tujuan untuk menyeragamkan bentuk teks agar representasi data yang dihasilkan lebih konsisten [19]. Dalam penelitian ini, proses normalisasi mencakup konversi emoji dan emotikon menjadi representasi teks, normalisasi karakter berulang, penghapusan kata atau frasa yang berulang, serta normalisasi spasi. Melalui tahapan tersebut, variasi penulisan yang tidak memengaruhi makna dapat diminimalkan sehingga kualitas data dapat ditingkatkan.

C. Persiapan Label

Pada tahap ini dilakukan persiapan label emosi pada dataset GoEmotions sebelum digunakan pada proses pelatihan model. Dataset GoEmotions merupakan dataset *multi-label*, yaitu sebuah sampel dapat memiliki beberapa label emosi. Namun, sekitar 83% sampel pada dataset GoEmotions hanya memiliki satu label emosi atau *single-label* [5]. Berdasarkan karakteristik tersebut, dilakukan penyaringan data dengan hanya mempertahankan sampel *single-label*, sedangkan sampel yang memiliki lebih dari satu label emosi tidak digunakan dalam penelitian ini.

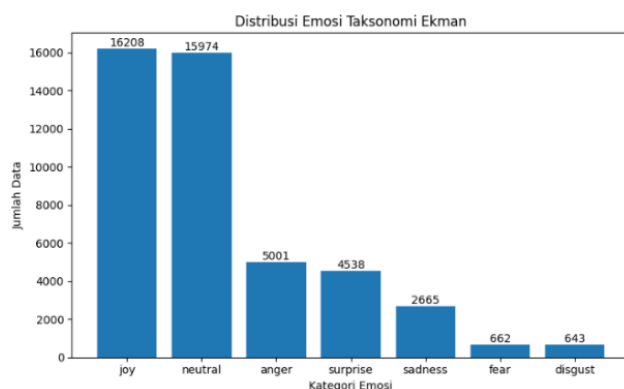
Selanjutnya dilakukan pemetaan label emosi berdasarkan skema hierarki taksonomi Ekman yang diperkenalkan oleh Demszy dkk. [5]. Berdasarkan skema tersebut, 27 kategori emosi pada GoEmotions dikelompokkan ke dalam enam emosi dasar, yaitu *anger*, *disgust*, *fear*, *joy*, *sadness*, dan *surprise*, sedangkan kategori *neutral* tetap dipertahankan. Penggunaan pengelompokan Ekman dipilih karena pengelompokan tersebut dapat mengurangi kebingungan antar kategori emosi yang lebih spesifik dalam kelompok yang sama [5]. Rincian pemetaan setiap kategori emosi ditunjukkan pada Tabel I. Selanjutnya, hasil pemetaan label dikonversi menjadi representasi numerik menggunakan *label*

encoding untuk memperoleh nilai label yang digunakan sebagai target dalam pelatihan model.

TABLE I
PEMETAAN LABEL EMOSI

Label GoEmotions	Label Ekman
anger, annoyance, disapproval	anger
disgust	disgust
fear, nervousness	fear
admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief	joy
disappointment, embarrassment, grief, remorse, sadness	sadness
confusion, curiosity, realization, surprise	surprise
neutral	neutral

Berdasarkan hasil pemetaan ke dalam taksonomi Ekman, diperoleh tujuh kategori emosi dengan distribusi data yang ditunjukkan pada gambar 2. Kategori *joy* dan *neutral* merupakan kelas dengan jumlah sampel terbanyak, sedangkan kategori *fear* dan *disgust* merupakan kelas dengan jumlah sampel terendah. Ketimpangan jumlah sampel antar kategori yang cukup besar mengindikasikan distribusi data dalam penelitian ini belum merata. Oleh karena itu, penelitian ini menerapkan penanganan *class imbalance* menggunakan *class weights* agar model tidak memberikan kecenderungan prediksi terhadap kelas mayoritas serta tetap mempertahankan kemampuan identifikasi terhadap kelas yang memiliki jumlah sampel lebih rendah.



Gambar 2. Distribusi Kelas

D. Tokenisasi

Pada tahap ini, tokenisasi diterapkan terhadap teks hasil prapemrosesan dengan *tokenizer* yang disesuaikan dengan arsitektur IndoRoBERTa dan EmoRoBERTa. Hasil tokenisasi kemudian direpresentasikan melalui *input_ids* dan *attention_mask* sebagai dua komponen masukan bagi model transformer [21]. Nilai *max_length* ditetapkan sebesar 128 token berdasarkan hasil analisis distribusi panjang token pada masing-masing model sehingga kedua model menggunakan panjang input yang seragam pada tahap pelatihan maupun pengujian.

Selanjutnya, hasil tokenisasi digunakan sebagai input pada proses *fine-tuning* masing-masing model.

E. Transformasi Data Menggunakan TF-IDF

Pada baseline konvensional, data teks hasil prapemrosesan ditransformasikan menjadi representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini memberikan bobot pada setiap term berdasarkan tingkat kepentingannya dalam suatu dokumen dan keseluruhan kumpulan dokumen [22]. Dalam penelitian ini, TF-IDF digunakan dengan mempertimbangkan kata tunggal dan pasangan kata yang berurutan. Term dengan frekuensi kemunculan yang terlalu rendah maupun terlalu tinggi disaring untuk menghasilkan fitur yang lebih relevan, sedangkan penskalaan logaritmik diterapkan pada frekuensi term. Hasil transformasi TF-IDF selanjutnya digunakan sebagai masukan pada Linear Support Vector Machine (SVM) untuk melakukan klasifikasi emosi.

F. Penanganan Data Imbalance

Ketidakmerataan jumlah sampel antar kategori emosi dalam dataset penelitian menyebabkan munculnya kondisi *class imbalance*, yaitu distribusi sampel yang berbeda pada masing-masing kelas. Kondisi ini ditangani dengan memberikan bobot yang berbeda terhadap setiap kelas melalui teknik *class weighting* pada proses *fine-tuning*. Penelitian sebelumnya menyebutkan bahwa *inverse of square root of number of samples* (ISNS) dapat digunakan sebagai pendekatan pembobotan kelas untuk mengurangi pengaruh ketimpangan distribusi data dan meningkatkan performa model pada *simulated datasets* [23]. Berdasarkan pendekatan tersebut, bobot untuk masing-masing kelas dalam penelitian ini dihitung menggunakan metode ISNS, kemudian diterapkan pada fungsi *CrossEntropyLoss* selama proses *fine-tuning*.

G. Hyperparameter Tuning

Tahap *hyperparameter tuning* dilakukan dengan menguji sejumlah konfigurasi parameter guna menentukan kombinasi yang menghasilkan performa optimal bagi model [20]. Pada penelitian ini, parameter yang digunakan dalam proses *hyperparameter tuning* disesuaikan dengan karakteristik masing-masing model. Pada IndoRoBERTa, EmoRoBERTa, dan mBERT, *learning rate* dan *batch size* menjadi parameter yang divariasikan, sedangkan parameter pelatihan lainnya dijaga tetap pada setiap skenario pengujian.

Sementara itu, pada *baseline* SVM yang menggunakan TF-IDF sebagai representasi fitur, *hyperparameter tuning* dilakukan dengan menguji beberapa nilai C sebagai parameter regularisasi pada Linear SVM. Nilai C yang diuji adalah 0.1, 1, dan 10, kemudian setiap konfigurasi dievaluasi berdasarkan nilai macro F1 pada data validasi. Pendekatan serupa juga diterapkan pada penelitian sebelumnya, yaitu dengan melakukan tuning parameter C pada SVM dan menggunakan F1-score untuk menentukan

konfigurasi yang optimal [24]. Hasil dari setiap konfigurasi kemudian dibandingkan untuk menentukan konfigurasi optimal yang digunakan pada tahap pelatihan dan evaluasi masing-masing model. Rangkaian konfigurasi *hyperparameter* untuk model Transformer disajikan pada Tabel II.

TABLE II
NILAI UJI HYPERPARAMETER PADA MODEL TRANSFORMER

Hyperparameter	Nilai
Learning Rate	1e-5, 2e-5
Batch Size	16, 32
Epoch	10
Max Length	128

H. Evaluasi

Evaluasi klasifikasi emosi dalam penelitian ini mencakup *accuracy*, *macro precision*, *macro recall*, dan *macro F1-score*. Penilaian utama performa model mengacu pada *macro F1-score*, sedangkan metrik lainnya digunakan sebagai pelengkap untuk menggambarkan hasil klasifikasi secara keseluruhan. Penggunaan *macro average* didasarkan pada kondisi distribusi kelas dalam dataset yang tidak merata, yaitu nilai evaluasi setiap kelas dihitung secara individual dan selanjutnya dirata-ratakan sehingga setiap kategori memperoleh kontribusi yang setara dalam perhitungan tanpa dipengaruhi oleh jumlah sampel pada masing-masing kelas, serta mampu memberikan representasi kinerja model secara lebih proporsional dan tepat untuk mengevaluasi performa model dalam kondisi distribusi kelas yang tidak merata [25].

Pada permasalahan klasifikasi multikelas yang terdiri atas C kelas, perhitungan metrik evaluasi didasarkan pada skema *one-vs-rest*. Pada skema ini, kelas yang sedang dievaluasi diposisikan sebagai kelas positif, sedangkan seluruh kategori lainnya diperlakukan sebagai kelas negatif. [26]. Berdasarkan pendekatan tersebut, untuk setiap kelas diperoleh empat komponen evaluasi, yaitu *True Positive* (TP_c), *False Positive* (FP_c), *False Negative* (FN_c), dan *True Negative* (TN_c). Nilai TP_c menunjukkan jumlah sampel dari kelas ke-c yang berhasil dikenali oleh model secara tepat, FP_c menunjukkan sampel dari kelas lain yang salah dikenali sebagai kelas ke-c, FN_c menunjukkan jumlah sampel kelas ke-c yang justru dikategorikan ke kelas lainnya, sedangkan TN_c menunjukkan sampel di luar kelas ke-c yang berhasil dikategorikan bukan sebagai kelas tersebut. Keempat nilai tersebut selanjutnya digunakan untuk memperoleh nilai *precision*, *recall*, *F1-score*, dan *accuracy* pada evaluasi performa model.

- 1) Metrik *precision* digunakan untuk mengukur tingkat ketepatan model dalam menghasilkan prediksi pada

suatu kelas. Nilai metrik ini diperoleh dengan membandingkan antara sampel yang berhasil diklasifikasikan secara benar sebagai kelas tersebut dan keseluruhan sampel yang diprediksi sebagai kelas tersebut [26]. Nilai *precision* dihitung menggunakan Persamaan (1).

$$\text{Precision} = \frac{TP_c}{TP_c + FP_c} \quad (1)$$

- 2) Metrik *recall* digunakan untuk menilai kemampuan model dalam menangkap seluruh sampel yang termasuk ke dalam suatu kelas. Nilai *recall* menggambarkan proporsi sampel dari kelas tersebut yang berhasil dikenali dengan benar dibandingkan dengan keseluruhan sampel yang sebenarnya termasuk dalam kelas tersebut. [26]. Nilai *recall* dihitung menggunakan Persamaan (2).

$$\text{Recall} = \frac{TP_c}{TP_c + FN_c} \quad (2)$$

- 3) Metrik *F1-score* digunakan untuk menggambarkan performa klasifikasi dengan menggabungkan nilai *precision* dan *recall* melalui perhitungan rata-rata harmonik. Metrik ini memberikan penilaian yang mempertimbangkan ketepatan prediksi sekaligus tingkat keberhasilan model dalam mengidentifikasi sampel yang termasuk ke dalam masing-masing kelas [26]. Nilai *F1-score* dihitung menggunakan Persamaan (3).

$$F - 1 \text{ Score} = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (3)$$

- 4) Metrik *macro precision* menunjukkan rata-rata ketepatan model dalam mengidentifikasi setiap kelas emosi dengan memperlakukan seluruh kelas secara setara. Nilai *macro precision* diperoleh dari nilai *precision* pada setiap kelas yang kemudian dirata-ratakan, sehingga hasil evaluasi tidak dipengaruhi oleh dominasi jumlah sampel pada kelas tertentu [27]. Perhitungan *macro precision* ditunjukkan pada Persamaan (4).

$$\text{Macro Precision} = \frac{1}{C} \sum_{c=1}^C \text{Precision}_c \quad (4)$$

- 5) Metrik *macro recall* menunjukkan rata-rata kemampuan model dalam mengenali sampel dari setiap kelas emosi. Nilai *macro recall* diperoleh dengan merata-ratakan nilai *recall* pada masing-masing kelas. Dengan cara tersebut, kemampuan model dalam mendeteksi setiap kelas memperoleh penilaian yang setara, tanpa menjadikan perbedaan jumlah sampel pada masing-masing kelas sebagai

faktor yang memengaruhi hasil metrik [27]. Perhitungan *macro recall* ditunjukkan pada Persamaan (5).

$$\text{Macro Recall} = \frac{1}{C} \sum_{c=1}^C \text{Recall}_c \quad (5)$$

- 6) Metrik *macro F1-score* digunakan untuk menggambarkan performa model dengan mempertimbangkan hasil klasifikasi pada seluruh kelas. Nilai metrik ini berasal dari rata-rata *F1-score* setiap kelas yang dihitung berdasarkan *precision* dan *recall* masing-masing kelas. Dengan pendekatan tersebut, evaluasi memberikan porsi yang setara bagi setiap kelas, termasuk ketika jumlah sampel antar kelas dalam dataset tidak merata. Atas pertimbangan tersebut, *macro F1-score* ditetapkan sebagai metrik utama dalam penelitian ini [27]. Perhitungan *macro F1-score* ditunjukkan pada Persamaan (6).

$$\text{Macro F1-score} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (6)$$

- 7) *Accuracy* menunjukkan proporsi seluruh sampel yang berhasil diklasifikasikan secara tepat dibandingkan dengan keseluruhan sampel pada data uji, sehingga performa model dapat dinilai secara menyeluruh [26]. Meskipun metrik ini dapat dipengaruhi oleh ketidakseimbangan distribusi kelas, *accuracy* juga diperhitungkan untuk melengkapi penilaian performa melalui metrik *macro average*. Perhitungan *accuracy* ditunjukkan pada Persamaan (7).

$$\text{Accuracy} = \frac{\sum_c TP_c}{N} \quad (7)$$

Selain itu, *confusion matrix* digunakan untuk mengevaluasi hasil klasifikasi pada setiap kelas emosi. Matriks ini memperlihatkan hubungan antara hasil prediksi model dan label sebenarnya melalui empat komponen, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Informasi tersebut memungkinkan kemampuan model dalam membedakan kategori emosi ditinjau secara lebih rinci [28]. Selain menggambarkan tingkat keberhasilan klasifikasi, analisis ini juga memungkinkan pola kesalahan prediksi yang dihasilkan model untuk diidentifikasi, sehingga hasil penilaian terhadap performa model dapat diperoleh secara lebih menyeluruh, termasuk pada kondisi dataset yang memiliki distribusi kelas tidak merata.

III. HASIL DAN PEMBAHASAN

Kinerja model dalam klasifikasi emosi dapat dilihat dari sejauh mana model mampu mengenali berbagai kategori emosi pada data yang diberikan. Pada penelitian ini, pengujian kinerja difokuskan pada perbandingan IndoRoBERTa dan EmoRoBERTa sebagai model utama dalam klasifikasi emosi pada teks berbahasa Indonesia, sedangkan mBERT dan SVM yang menggunakan TF-IDF digunakan sebagai *baseline* pembanding untuk memberikan konteks terhadap kinerja kedua model utama. Data pengujian bersumber dari dataset GoEmotions terjemahan bahasa Indonesia yang telah diseleksi menjadi data *single-label* dan dipetakan ke dalam tujuh kategori emosi berdasarkan taksonomi Ekman. Dataset yang digunakan terdiri atas 45.691 data, yang terbagi menjadi 36.572 data latih, 4.538 data validasi, dan 4.581 data uji.

Kinerja masing-masing model selanjutnya dievaluasi menggunakan *accuracy*, *macro precision*, *macro recall*, dan *macro F1-score*, dengan *macro F1-score* sebagai metrik utama. Analisis diarahkan pada perbandingan kinerja IndoRoBERTa dan EmoRoBERTa, sedangkan hasil mBERT dan SVM dengan TF-IDF digunakan sebagai konteks tambahan untuk melihat capaian model utama terhadap pendekatan *multilingual Transformer* dan pendekatan *machine learning* konvensional.

A. Hyperparameter Tuning

Kinerja model dalam melakukan klasifikasi tidak terlepas dari konfigurasi parameter yang digunakan selama proses pelatihan. Berdasarkan hasil tersebut, proses *hyperparameter tuning* dilakukan untuk menentukan konfigurasi yang optimal bagi masing-masing model. Parameter yang diuji disesuaikan dengan karakteristik setiap pendekatan. Pada IndoRoBERTa, EmoRoBERTa, dan mBERT, *hyperparameter tuning* dilakukan dengan menguji kombinasi *learning rate* dan *batch size*, sedangkan parameter pelatihan lainnya dijaga tetap. Nilai *learning rate* yang diuji adalah 1×10^{-5} dan 2×10^{-5} , sedangkan *batch size* yang digunakan adalah 16 dan 32, sehingga diperoleh empat kombinasi *hyperparameter* untuk masing-masing model Transformer. Nilai *max length* ditetapkan sebesar 128 token berdasarkan hasil analisis distribusi panjang token dan digunakan secara konsisten pada tahap pelatihan maupun pengujian.

Sementara itu, pada *baseline SVM* dengan TF-IDF, *hyperparameter tuning* dilakukan dengan menguji nilai *C* sebagai parameter regularisasi pada Linear SVM. Setiap konfigurasi dievaluasi berdasarkan nilai *macro F1-score* pada data validasi. Seluruh model menggunakan mekanisme *class weighting* sesuai dengan pendekatan yang diterapkan pada masing-masing model, sedangkan proses pelatihan model Transformer dibatasi hingga 10 *epoch* dan dilengkapi dengan mekanisme *early stopping*. Konfigurasi terbaik ditentukan berdasarkan nilai *macro F1-score* dan digunakan pada tahap pengujian. Hasil konfigurasi terbaik masing-masing model disajikan pada Tabel III.

TABLE III
HASIL KONFIGURASI TERBAIK
PADA MASING-MASING MODEL

Model	Konfigurasi Terbaik
IndoRoBERTa	Learning Rate: 1×10^{-5} ; Batch Size: 16
EmoRoBERTa	Learning Rate: 2×10^{-5} ; Batch Size: 16
mBERT	Learning Rate: 2×10^{-5} ; Batch Size: 16
SVM + TF-IDF	C = 1

Berdasarkan hasil *hyperparameter tuning* yang disajikan pada Tabel III, setiap model memperoleh konfigurasi yang berbeda sesuai dengan karakteristik pendekatan yang digunakan setelah seluruh percobaan dilakukan. Pada kelompok model Transformer, IndoRoBERTa memperoleh konfigurasi terbaik dengan *learning rate* sebesar 1×10^{-5} dan batch size 16, sedangkan EmoRoBERTa dan mBERT memperoleh konfigurasi terbaik dengan *learning rate* sebesar 2×10^{-5} dan batch size 16. Perbedaan konfigurasi tersebut menunjukkan bahwa kombinasi parameter yang memberikan hasil validasi terbaik tidak sepenuhnya sama pada setiap model, meskipun pengujian dilakukan pada rentang parameter yang sama.

Pada pendekatan konvensional, SVM dengan TF-IDF memperoleh nilai C=1 sebagai konfigurasi terbaik. Berbeda dengan model Transformer yang menggunakan *learning rate* dan *batch size* sebagai parameter yang diuji, proses tuning pada SVM difokuskan pada parameter regularisasi C. Dengan demikian, pemilihan konfigurasi pada masing-masing model dilakukan berdasarkan parameter yang relevan terhadap karakteristik metode yang digunakan. Konfigurasi terpilih selanjutnya digunakan dalam proses pelatihan dan pengujian untuk memperoleh kinerja akhir masing-masing model.

TABLE IV
HASIL EVALUASI KONFIGURASI TERBAIK PADA MASING-MASING MODEL

Metrik Evaluasi	IndoRoBERTa	EmoRoBERTa	mBERT	SVM
Accuracy	0.6603	0.6326	0.6195	0.5999
Macro Precision	0.5938	0.5613	0.5543	0.5362
Macro Recall	0.5975	0.5956	0.5504	0.5640
Macro F-1 Score	0.5887	0.5742	0.5495	0.5449

Berdasarkan hasil evaluasi pada Tabel IV, penggunaan model dengan karakteristik representasi yang berbeda menghasilkan variasi capaian pada tugas klasifikasi emosi. IndoRoBERTa menghasilkan *accuracy* sebesar 0.6603, *macro precision* sebesar 0.5938, *macro recall* sebesar 0.5975, dan *macro F1-score* sebesar 0.5887. Sementara itu, EmoRoBERTa memperoleh *accuracy* sebesar 0.6326, *macro precision* sebesar 0.5613, *macro recall* sebesar 0.5956, dan *macro F1-score* sebesar 0.5742. Perbedaan tersebut menunjukkan adanya variasi kinerja antara kedua model

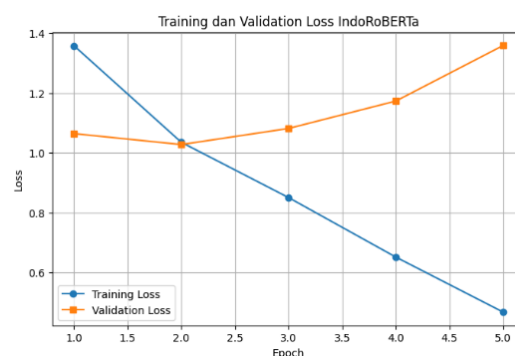
utama ketika diterapkan pada dataset GoEmotions versi bahasa Indonesia dengan skema klasifikasi dan pembagian data yang sama.

Selain itu, untuk memberikan konteks terhadap capaian kedua model utama, penelitian ini turut menggunakan mBERT dan SVM dengan TF-IDF sebagai *baseline*. mBERT merepresentasikan pendekatan multilingual Transformer yang tidak secara khusus dikembangkan untuk bahasa Indonesia maupun klasifikasi emosi, sedangkan SVM dengan TF-IDF merepresentasikan pendekatan konvensional yang mengandalkan representasi berbasis frekuensi term dan klasifikasi berbasis margin. Dengan menempatkan kedua pendekatan tersebut pada kondisi evaluasi yang sama, kinerja model utama dapat ditinjau tidak hanya berdasarkan perbedaan antara dua model dengan orientasi spesialisasi yang berbeda, tetapi juga dalam konteks pendekatan multilingual dan pendekatan berbasis fitur konvensional.

Hasil evaluasi menunjukkan bahwa mBERT memperoleh *accuracy* sebesar 0.6195, *macro precision* sebesar 0.5543, *macro recall* sebesar 0.5504, dan *macro F1-score* sebesar 0.5495. Sementara itu, SVM + TF-IDF memperoleh *accuracy* sebesar 0.5999, *macro precision* sebesar 0.5362, *macro recall* sebesar 0.5640, dan *macro F1-score* sebesar 0.5449. Nilai tersebut memberikan titik pembandingan empiris terhadap hasil IndoRoBERTa dan EmoRoBERTa pada eksperimen yang sama. Dengan demikian, penggunaan mBERT dan SVM menggunakan TF-IDF sebagai *baseline* memberikan dasar pembandingan yang memungkinkan capaian IndoRoBERTa dan EmoRoBERTa diinterpretasikan secara lebih kontekstual. Evaluasi terhadap pendekatan multilingual dan konvensional pada dataset serta skema klasifikasi yang sama membantu menunjukkan apakah capaian kedua model utama tetap terlihat ketika dibandingkan dengan pendekatan yang memiliki karakteristik representasi yang berbeda.

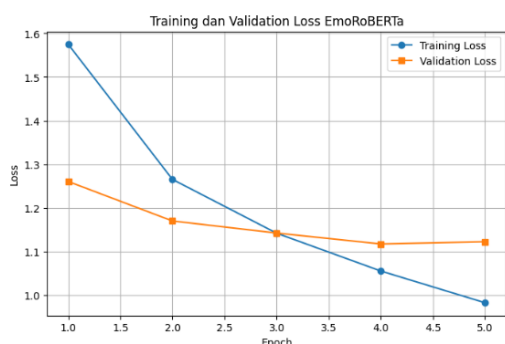
B. Pelatihan Model

Setelah konfigurasi terbaik diperoleh melalui tahap *hyperparameter tuning*, proses pembelajaran pada dua model utama, yaitu IndoRoBERTa dan EmoRoBERTa, dianalisis melalui perkembangan *training loss* dan *validation loss* pada setiap *epoch*.



Gambar 3. Training dan Validation Loss IndoRoBERTa

Analisis ini digunakan untuk mengamati dinamika pembelajaran model pada data latih dan validasi, serta mengidentifikasi indikasi perubahan kemampuan generalisasi selama proses *fine-tuning*. Berdasarkan pola *loss* yang ditampilkan pada gambar 3, IndoRoBERTa memperlihatkan penurunan *training loss* secara konsisten sepanjang *epoch*. Di sisi lain, *validation loss* menurun pada beberapa *epoch* awal sebelum mengalami peningkatan pada *epoch* berikutnya. Perbedaan pola tersebut mengindikasikan adanya gejala *overfitting ringan*, yang ditunjukkan oleh kondisi ketika penurunan *training loss* tidak lagi diikuti oleh penurunan *validation loss*. Hal ini menunjukkan bahwa model semakin menyesuaikan pembelajarannya terhadap karakteristik data latih, sementara kemampuan generalisasi terhadap data validasi mulai mengalami penurunan.



Gambar 4. Training dan Validation Loss EmoRoBERTa

Berdasarkan pola *loss* pada Gambar 4, EmoRoBERTa memperlihatkan penurunan *training loss* dan *validation loss* sepanjang proses pelatihan. Penurunan *training loss* mengindikasikan bahwa model secara bertahap mempelajari pola yang terdapat pada data latih, sementara penurunan *validation loss* mencerminkan peningkatan kemampuan model dalam menghasilkan prediksi terhadap data validasi. Keselarasan antara kedua nilai *loss* tersebut menunjukkan bahwa proses pembelajaran EmoRoBERTa berlangsung relatif stabil selama pelatihan. Selain itu, *validation loss* tidak mengalami peningkatan secara konsisten meskipun *training loss* terus menurun. Kondisi tersebut mengindikasikan bahwa EmoRoBERTa tidak memperlihatkan gejala *overfitting* yang menonjol selama proses pelatihan.

C. Pengujian Model

Setelah konfigurasi terbaik diperoleh melalui tahap *hyperparameter tuning*, seluruh pendekatan yang digunakan dalam penelitian ini selanjutnya dievaluasi menggunakan data uji untuk mengukur kinerja klasifikasi pada tujuh kategori emosi. Evaluasi dilakukan pada tingkat kategori dengan mempertimbangkan nilai *precision*, *recall*, dan *F1-score*, sedangkan *accuracy* digunakan sebagai metrik tambahan untuk memberikan gambaran kinerja klasifikasi secara keseluruhan. Analisis pada tingkat kategori dilakukan untuk mengidentifikasi kemampuan masing-masing pendekatan dalam mengenali setiap kategori emosi serta melihat pola

kinerja pada kategori dengan jumlah sampel yang berbeda. Hasil evaluasi untuk setiap kategori emosi disajikan pada Tabel V.

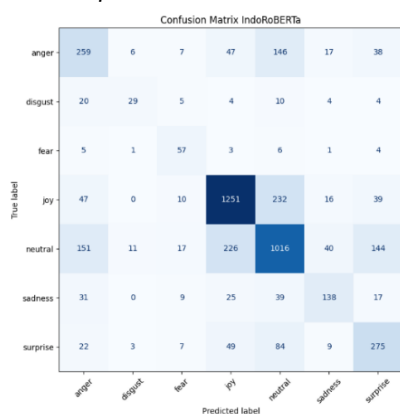
TABLE V
HASIL PENGUJIAN SETIAP KATEGORI EMOSI

Model	Class	Accuracy	Precision	Recall	F1-score
IndoRoBERTa	Anger (0)	0,6603	0.48	0.50	0.49
	Disgust (1)		0.58	0.38	0.46
	Fear (2)		0.51	0.74	0.60
	Joy (3)		0.78	0.78	0.78
	Neutral (4)		0.66	0.63	0.65
	Sadness (5)		0.61	0.53	0.57
	Surprise (6)		0.53	0.61	0.57
EmoRoBERTa	Anger (0)	0.6326	0.48	0.44	0.46
	Disgust (1)		0.49	0.46	0.48
	Fear (2)		0.60	0.68	0.64
	Joy (3)		0.78	0.74	0.76
	Neutral (4)		0.64	0.59	0.62
	Sadness (5)		0.44	0.63	0.51
	Surprise (6)		0.49	0.63	0.55
mBERT	Anger (0)	0.6195	0.40	0.46	0.42
	Disgust (1)		0.46	0.36	0.40
	Fear (2)		0.59	0.61	0.60
	Joy (3)		0.73	0.79	0.76
	Neutral (4)		0.65	0.54	0.59
	Sadness (5)		0.57	0.52	0.54
	Surprise (6)		0.49	0.58	0.53
SVM + TF-IDF	Anger (0)	0.5999	0.37	0.50	0.43
	Disgust (1)		0.44	0.38	0.41
	Fear (2)		0.62	0.69	0.65
	Joy (3)		0.82	0.66	0.73
	Neutral (4)		0.62	0.60	0.61
	Sadness (5)		0.45	0.56	0.50
	Surprise (6)		0.43	0.55	0.49

Berdasarkan hasil pengujian pada Tabel V, kategori *joy* memperoleh *F1-score* tertinggi pada seluruh model, yaitu

sebesar 0,78 pada IndoRoBERTa, 0,76 pada EmoRoBERTa, 0,76 pada mBERT, dan 0,73 pada SVM dengan TF-IDF. Sebaliknya, capaian terendah pada masing-masing model terdapat pada kategori yang berbeda. IndoRoBERTa menghasilkan *F1-score* sebesar 0,46 pada kategori *disgust*, EmoRoBERTa sebesar 0,46 pada kategori *anger*, mBERT sebesar 0,40 pada kategori *disgust*, dan SVM dengan TF-IDF sebesar 0,41 pada kategori *disgust*.

Perbedaan capaian pada masing-masing kategori menunjukkan bahwa performa klasifikasi tidak terdistribusi secara merata pada seluruh kelas. Kondisi tersebut berkaitan dengan distribusi data yang tidak seimbang, sehingga jumlah sampel yang tersedia pada setiap kategori berbeda. Kategori dengan jumlah data yang lebih sedikit memiliki jumlah contoh yang lebih terbatas untuk digunakan dalam proses pembelajaran. Meskipun *class weighting* telah diterapkan selama proses pelatihan untuk mengakomodasi ketidakseimbangan kelas, perbedaan capaian antarkategori masih terlihat pada hasil pengujian. Untuk mengetahui lebih lanjut pola prediksi yang membentuk capaian tersebut, distribusi hasil klasifikasi pada masing-masing model dianalisis melalui *confusion matrix*.

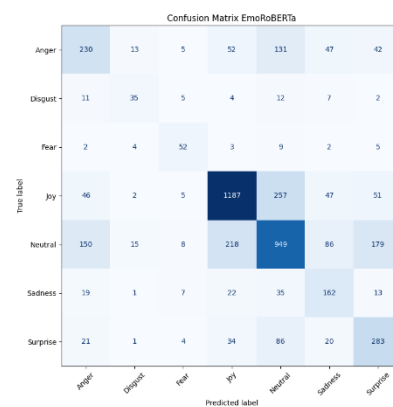


Gambar 5. *Confusion Matrix* IndoRoBERTa

Distribusi hasil prediksi IndoRoBERTa pada tujuh kategori emosi ditampilkan melalui *confusion matrix* pada Gambar 5. Prediksi pada diagonal menunjukkan bahwa kategori *joy* dan *neutral* memiliki jumlah klasifikasi benar yang paling besar, masing-masing sebanyak 1.251 dan 1.016 data. Sebaliknya, jumlah prediksi benar pada *disgust* dan *fear* relatif lebih rendah, yaitu masing-masing 29 dan 57 data. Pola tersebut memperlihatkan bahwa hasil klasifikasi tidak terdistribusi secara merata pada seluruh kategori, sebagaimana tercermin pula pada perbedaan capaian metrik evaluasi antarkategori.

Selain prediksi yang tepat, *confusion matrix* memperlihatkan adanya kecenderungan kesalahan klasifikasi pada kategori tertentu. Pertukaran prediksi antara *joy* dan *neutral* merupakan pola yang paling menonjol, dengan 232 data aktual *joy* diprediksi sebagai *neutral* dan 226 data aktual *neutral* diprediksi sebagai *joy*. Kesalahan klasifikasi juga terlihat pada kategori *anger* dan *surprise* yang masing-masing mengalami 146 dan 144 prediksi *neutral*. Dengan demikian, kesalahan klasifikasi IndoRoBERTa tidak tersebar secara

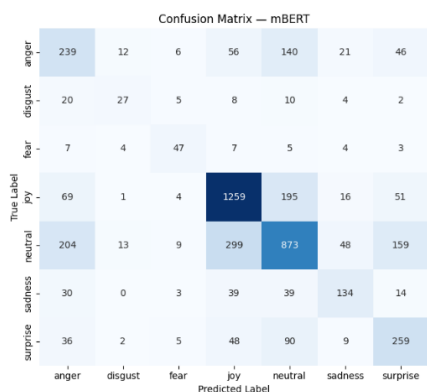
acak, melainkan terkonsentrasi pada sejumlah pasangan kategori tertentu, terutama yang melibatkan *joy* dan *neutral*. Pola ini memberikan penjelasan terhadap variasi capaian metrik pada tingkat kategori yang telah ditunjukkan pada Tabel V.



Gambar 6. *Confusion Matrix* EmoRoBERTa

Distribusi hasil klasifikasi EmoRoBERTa menunjukkan kecenderungan yang serupa pada beberapa kategori emosi, sebagaimana ditampilkan melalui *confusion matrix* pada Gambar 6. Prediksi benar paling banyak terdapat pada *joy* dan *neutral*, masing-masing sebanyak 1.187 dan 949 data, sedangkan *disgust* dan *fear* memiliki prediksi benar yang lebih rendah, yaitu 35 dan 52 data. Temuan tersebut menunjukkan bahwa distribusi prediksi benar pada EmoRoBERTa juga terkonsentrasi pada kategori tertentu dan sejalan dengan variasi capaian evaluasi yang ditunjukkan pada Tabel V.

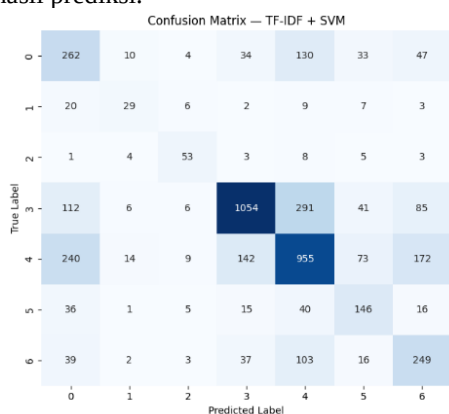
Pola kesalahan klasifikasi kembali menunjukkan keterkaitan yang kuat antara *joy* dan *neutral*. Sebanyak 257 data aktual *joy* diklasifikasikan sebagai *neutral*, sedangkan 218 data aktual *neutral* diklasifikasikan sebagai *joy*. Selain itu, 150 data *anger* diprediksi sebagai *neutral* dan 179 data *neutral* diprediksi sebagai *surprise*. Pada *sadness*, terdapat 162 prediksi yang tepat, sementara kesalahan klasifikasi terutama mengarah ke *joy* dan *anger*. Pola tersebut menunjukkan bahwa kesalahan prediksi EmoRoBERTa juga terkonsentrasi pada kategori tertentu, dengan *joy* dan *neutral* menjadi pasangan yang paling dominan dalam pertukaran prediksi. Perbedaan pola kesalahan antara kedua model selanjutnya menjadi salah satu aspek yang dipertimbangkan dalam perbandingan kinerja IndoRoBERTa dan EmoRoBERTa.



Gambar 7. Confusion Matrix Mbert

Distribusi hasil klasifikasi mBERT menunjukkan bahwa prediksi benar paling banyak terdapat pada kategori *joy* dan *neutral*, masing-masing sebanyak 1.259 dan 873 data, sebagaimana ditunjukkan pada *confusion matrix* pada Gambar 7. Sementara itu, jumlah prediksi benar yang lebih rendah ditemukan pada kategori *disgust* dan *fear*, yaitu sebanyak 27 dan 47 data. Temuan tersebut menunjukkan bahwa distribusi prediksi benar pada mBERT juga terkonsentrasi pada kategori tertentu dan sejalan dengan variasi capaian evaluasi yang ditunjukkan pada Tabel V.

Pola kesalahan klasifikasi menunjukkan bahwa pertukaran prediksi paling dominan terjadi antara kategori *joy* dan *neutral*. Sebanyak 195 data aktual *joy* diprediksi sebagai *neutral*, sedangkan 299 data aktual *neutral* diprediksi sebagai *joy*. Selain itu, 140 data *anger* diprediksi sebagai *neutral*, sementara 204 data *neutral* diklasifikasikan sebagai *anger*. Pada kategori *sadness*, terdapat 134 prediksi yang tepat, sedangkan kesalahan klasifikasi paling banyak mengarah ke *joy* dan *neutral*, masing-masing sebanyak 39 data. Pola tersebut menunjukkan bahwa kesalahan prediksi mBERT masih didominasi oleh keterkaitan antara kategori *joy* dan *neutral*, dengan kategori *neutral* menjadi salah satu sumber utama kesalahan klasifikasi baik sebagai kelas aktual maupun sebagai hasil prediksi.



Gambar 8. Confusion Matrix SVM Menggunakan TF-IDF

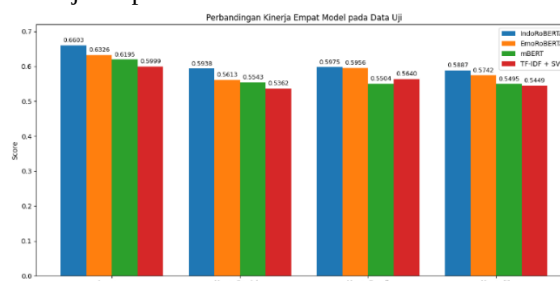
Distribusi hasil klasifikasi SVM dengan TF-IDF menunjukkan bahwa prediksi benar paling banyak terdapat pada kategori *joy* dan *neutral*, masing-masing sebanyak 1.054

dan 955 data, sebagaimana ditunjukkan pada *confusion matrix* pada Gambar 8. Sementara itu, jumlah prediksi benar yang lebih rendah ditemukan pada kategori *disgust* dan *fear*, yaitu sebanyak 29 dan 53 data. Temuan tersebut menunjukkan bahwa distribusi prediksi benar pada SVM dengan TF-IDF juga terkonsentrasi pada kategori tertentu dan sejalan dengan variasi capaian evaluasi yang ditunjukkan pada Tabel V.

Pola kesalahan klasifikasi menunjukkan bahwa pertukaran prediksi paling dominan terjadi antara kategori *joy* dan *neutral*. Sebanyak 291 data aktual *joy* diprediksi sebagai *neutral*, sedangkan 142 data aktual *neutral* diklasifikasikan sebagai *joy*. Selain itu, 130 data *anger* diprediksi sebagai *neutral*, sementara 240 data *neutral* diprediksi sebagai *anger*. Pada kategori *sadness*, terdapat 146 prediksi yang tepat, sedangkan kesalahan klasifikasi paling banyak mengarah ke *neutral* sebanyak 40 data. Sementara itu, kategori *surprise* menunjukkan 249 prediksi yang tepat dengan kesalahan klasifikasi terbesar mengarah ke *neutral* sebanyak 103 data. Pola tersebut menunjukkan bahwa kesalahan prediksi SVM dengan TF-IDF masih didominasi oleh keterkaitan antara kategori *joy* dan *neutral*, serta kecenderungan kategori *neutral* untuk tertukar dengan beberapa kategori emosi lainnya, terutama *anger* dan *surprise*.

D. Perbandingan Kinerja Model

Hasil evaluasi pada tingkat kategori menunjukkan bahwa capaian klasifikasi tidak terdistribusi secara merata pada seluruh emosi. Beberapa kategori, seperti *joy*, menunjukkan performa yang relatif tinggi pada seluruh model, sedangkan kategori lain masih menghasilkan capaian yang lebih rendah. Temuan tersebut menunjukkan bahwa kinerja model belum dapat direpresentasikan secara menyeluruh melalui hasil klasifikasi pada masing-masing kategori. Dengan demikian, evaluasi pada tingkat model secara keseluruhan diperlukan guna memperoleh pemahaman yang lebih menyeluruh terkait capaian klasifikasi emosi yang dihasilkan oleh setiap model. Perbandingan kinerja IndoRoBERTa, EmoRoBERTa, mBERT, serta SVM dengan TF-IDF berdasarkan metrik *accuracy*, *macro precision*, *macro recall*, dan *macro F1-score* disajikan pada Gambar 9.



Gambar 9. Perbandingan Kinerja Model

Berdasarkan Gambar 9, IndoRoBERTa memperoleh capaian tertinggi pada seluruh metrik evaluasi yang digunakan. Model tersebut mencatat *accuracy* sebesar

0.6603, *macro precision* sebesar 0.5938, *macro recall* sebesar 0.5975, dan *macro F1-score* sebesar 0.5887. EmoRoBERTa menempati posisi berikutnya dengan *accuracy* 0.6326, *macro precision* 0.5613, *macro recall* 0.5956, dan *macro F1-score* 0.5742. Sementara itu, mBERT dan SVM dengan TF-IDF yang digunakan sebagai *baseline* menghasilkan capaian yang relatif lebih rendah pada seluruh metrik evaluasi.

Sebagai metrik utama dalam penelitian ini, *macro F1-score* digunakan untuk mengevaluasi performa model secara menyeluruh karena mempertimbangkan keseimbangan antara *precision* dan *recall* pada seluruh kategori emosi. Berdasarkan metrik tersebut, IndoRoBERTa memperoleh nilai tertinggi sebesar 0.5887, diikuti oleh EmoRoBERTa sebesar 0.5742, mBERT sebesar 0.5495, dan SVM dengan TF-IDF sebesar 0.5449. Hasil tersebut menunjukkan bahwa, pada dataset dan pengaturan eksperimen yang digunakan dalam penelitian ini, kedua model utama menghasilkan capaian yang lebih tinggi dibandingkan model *baseline*.

Temuan tersebut turut didukung oleh capaian pada metrik evaluasi lainnya. Pada metrik *accuracy*, IndoRoBERTa dan EmoRoBERTa masing-masing memperoleh nilai 0.6603 dan 0.6326, sedangkan mBERT dan SVM dengan TF-IDF memperoleh nilai masing-masing sebesar 0.6195 dan 0.5999. Pola yang serupa juga terlihat pada *macro precision* dan *macro recall*, di mana kedua model utama secara konsisten mencatat nilai yang lebih tinggi dibandingkan model *baseline*. Konsistensi capaian tersebut menunjukkan bahwa perbedaan performa tidak hanya muncul pada satu metrik tertentu, tetapi juga tercermin pada berbagai aspek evaluasi yang digunakan dalam penelitian ini.

Meskipun demikian, selisih kinerja antara IndoRoBERTa dan EmoRoBERTa relatif lebih kecil dibandingkan selisih keduanya terhadap model *baseline*. Temuan ini menunjukkan bahwa kedua model memiliki karakteristik performa yang relatif berdekatan dalam melakukan klasifikasi emosi pada dataset yang digunakan, meskipun IndoRoBERTa secara konsisten menghasilkan nilai yang lebih tinggi pada seluruh metrik evaluasi. Di sisi lain, keberadaan mBERT dan SVM dengan TF-IDF sebagai *baseline* memberikan konteks tambahan dalam menginterpretasikan hasil penelitian. Berdasarkan pengujian yang dilakukan, kedua model utama memperoleh capaian yang lebih tinggi dibandingkan pendekatan multilingual maupun pendekatan konvensional yang digunakan sebagai pembanding.

Meskipun kinerja model yang diperoleh masih menunjukkan ruang untuk pengembangan, hasil penelitian ini memperlihatkan peningkatan capaian dibandingkan penelitian sebelumnya yang menggunakan dataset GoEmotions berbahasa Indonesia. Penelitian sebelumnya menerapkan IndoBERT dan DistilBERT pada 28 kategori emosi dalam skema klasifikasi multi-label, dengan capaian *macro F1-score* sebesar 0.4659 pada IndoBERT dan 0.4414 pada DistilBERT [12]. Sementara itu, penelitian ini menggunakan IndoRoBERTa dan EmoRoBERTa pada tujuh kategori emosi dengan skema klasifikasi yang disesuaikan dengan taksonomi Ekman, dan memperoleh *macro F1-score*

sebesar 0.5887 dan 0.5742 untuk masing-masing model. Hasil tersebut menunjukkan adanya peningkatan capaian *F1-score* pada klasifikasi emosi menggunakan GoEmotions berbahasa Indonesia, sekaligus menunjukkan potensi penggunaan model yang memiliki karakteristik berbeda dalam mendukung klasifikasi emosi pada bahasa Indonesia.

IV. KESIMPULAN

Hasil penelitian menunjukkan bahwa pemilihan model berperan dalam pencapaian klasifikasi emosi pada teks berbahasa Indonesia, khususnya ketika data memiliki distribusi kelas yang tidak seimbang. Penelitian ini menggunakan dataset GoEmotions versi bahasa Indonesia yang telah dipetakan ke dalam tujuh kategori emosi berdasarkan taksonomi Ekman, yaitu *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, dan *neutral*. Evaluasi dilakukan terhadap dua model utama, yaitu IndoRoBERTa dan EmoRoBERTa, dengan mBERT serta SVM dengan TF-IDF digunakan sebagai model *baseline*.

Berdasarkan hasil pengujian pada data uji, IndoRoBERTa memperoleh capaian tertinggi pada seluruh metrik evaluasi yang digunakan, dengan *accuracy* sebesar 0.6603, *macro precision* sebesar 0.5938, *macro recall* sebesar 0.5975, dan *macro F1-score* sebesar 0.5887. Sementara itu, EmoRoBERTa memperoleh *macro F1-score* sebesar 0.5742, diikuti oleh mBERT sebesar 0.5495 dan SVM dengan TF-IDF sebesar 0.5449. Hasil tersebut menunjukkan bahwa pada dataset dan pengaturan eksperimen yang digunakan dalam penelitian ini, kedua model utama menghasilkan capaian yang lebih tinggi dibandingkan model *baseline*.

Analisis pada tingkat kategori emosi menunjukkan bahwa performa klasifikasi tidak terdistribusi secara merata pada seluruh kelas. Kategori *joy* dan *neutral* secara umum memperoleh capaian yang lebih tinggi dibandingkan kategori lainnya, sedangkan kategori seperti *disgust* dan *fear* masih menunjukkan performa yang relatif lebih rendah. Hasil *confusion matrix* juga menunjukkan bahwa kesalahan klasifikasi masih banyak terjadi pada pasangan kategori tertentu, terutama antara *joy* dan *neutral*, yang mengindikasikan adanya kecenderungan pertukaran prediksi antara kedua kategori tersebut.

Secara keseluruhan, evaluasi yang dilakukan memperlihatkan bahwa pada dataset dan skenario klasifikasi yang digunakan, IndoRoBERTa menghasilkan performa yang lebih tinggi dibandingkan EmoRoBERTa. Selain itu, keberadaan mBERT dan SVM dengan TF-IDF sebagai *baseline* memberikan konteks tambahan dalam menginterpretasikan capaian kedua model utama. Meskipun demikian, performa klasifikasi yang diperoleh masih menunjukkan ruang untuk pengembangan lebih lanjut. Penelitian selanjutnya dapat mempertimbangkan penggunaan dataset yang lebih besar dan lebih seimbang, eksplorasi strategi penanganan ketidakseimbangan kelas yang berbeda, serta pengujian pada sumber data berbahasa Indonesia yang lebih beragam untuk mengevaluasi konsistensi performa model pada berbagai kondisi.

DAFTAR PUSTAKA

- [1] 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.
- [2] U. Yadav, P. Dasarwar, and D. Asudani, "Integrating commonsense knowledge with GPT embeddings for emotion classification," *MethodsX*, vol. 15, Dec. 2025, doi: 10.1016/j.mex.2025.103656.
- [3] E. Saraswati and Muljono, "Deteksi Emosi Pada Twitter Berbasis Fasttext: Evaluasi Performa Arsitektur CNN dan GRU," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 11, no. 1, pp. 1175–1186, Jan. 2026, doi: 10.36341/rabit.v11i1.7252.
- [4] S. R. Saputro, D. Arifianto, and T. A. Cahyanto, "Klasifikasi Emosi Pada Data Teks Pidato Politik Menggunakan Metode RoBERTa Emotion Classification In Political Speech Texts Using The ROBERTa Model," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 13, no. 1, pp. 211–220, Feb. 2026, doi: <https://doi.org/10.25126/jtiik.2026131>.
- [5] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2020, pp. 4040–4054. doi: 10.18653/v1/2020.acl-main.372.
- [6] R. Kamath, A. Ghoshal, S. Eswaran, and P. Honnavalli, "An Enhanced Context-based Emotion Detection Model using RoBERTa," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/CONECCT55679.2022.9865796.
- [7] L. Lomasto, N. Lettieri, D. Malandrino, V. Mosca, and R. Zaccagnino, "Modeling emotional signatures to detect conspiratorial communities on social media," *Neural Comput. Appl.*, vol. 38, no. 8, Apr. 2026, doi: 10.1007/s00521-026-12038-7.
- [8] C. M. Greco and A. Tagarelli, "Bringing order into the realm of Transformer-based language models for artificial intelligence and law," Dec. 01, 2024, *Springer Nature*. doi: 10.1007/s10506-023-09374-7.
- [9] E. Hassan, A. S. Talaat, and M. A. Elsabagh, "Intelligent text similarity assessment using Roberta with integrated chaotic perturbation optimization techniques," *J. Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-025-01233-3.
- [10] A. Araabi, C. Monz, and V. Niculae, "How Effective is Byte Pair Encoding for Out-Of-Vocabulary Words in Neural Machine Translation?" [Online]. Available: <https://github.com/pytorch/fairseq>
- [11] W. Suwarningsih, R. A. Pratama, F. Y. Rahadika, and M. H. A. Purnomo, "RoBERTa: language modelling in building Indonesian question-answering systems," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 6, pp. 1248–1255, Dec. 2022, doi: 10.12928/TELKOMNIKA.v20i6.24248.
- [12] J. R. Putri, Ermatita, and Abdiansah, "Class-Level Behavior Analysis under Metric Disagreement in Imbalanced Multi-Label Indonesian Emotion Classification," *Journal of Information Systems and Informatics*, vol. 8, no. 3, pp. 3595–3622, Jun. 2026, doi: 10.63158/journalisi.v8i3.1664.
- [13] M. Dzakwan Bintang Lazuardi and T. Noor Fatyanosa, "Klasifikasi Emosi Multikelas Berbasis Teks Bahasa Indonesia Menggunakan IndoRoBERTa," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 9, pp. 2548–964, Sep. 2025, doi: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/15276/6787>.
- [14] F. Fitriana and H. Setiawan, "Performance Analysis of SVM In Emotion Classification: A Comparative Study Of TF-IDF and Countvectorizer," *Journal of Embedded Systems, Security and Intelligent Systems*, pp. 133–145, Jun. 2025, doi: 10.59562/jessi.v6i2.8396.
- [15] A. Zamsuri, S. Defit, and G. W. Nurcahyo, "Development and Comparison of Multiple Emotion Classification Models in Indonesia Text Using Machine Learning," *Journal of Advances in Information Technology*, vol. 15, no. 4, pp. 519–531, 2024, doi: 10.12720/jait.15.4.519-531.
- [16] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?" [Online]. Available: <https://github.com/google-research/bert>
- [17] Y. O. Sihombing, K. Muchlisin, T. F. Arya, Moh. J. Mubarak, and R. F. Rachmadi, "A Systematic Evaluation of BERT Classifiers for Indonesia-based Text Data," *Techno.Com*, vol. 25, no. 2, pp. 232–243, May 2026, doi: 10.62411/tc.v25i2.15843.
- [18] M. Nurhadi, "Klasifikasi Emosi Multi-Label Pada Teks Berbahasa Indonesia Dengan Fine-Tuning Indobert Menggunakan Dataset GoEmotions," Sriwijaya University, 2025.
- [19] S. Alfariani, O. D. Yanti, S. N. M. Celesta, K. D. Tania, and A. Rifai, "Comparative Analysis of the Performance of Machine Learning Methods and Text Embedding Techniques in Classifying Toxic Conversations in the Roblox Game," *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 2775–2789, Jun. 2026, doi: 10.30871/jaic.v10i3.12646.
- [20] E. M. Pusung and I. N. Dewi, "Optimasi RoBERTa dengan Hyperparameter Tuning untuk Deteksi Emosi berbasis Teks," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 10, no. 3, pp. 240–248, Feb. 2025, doi: 10.25077/teknosi.v10i3.2024.240-248.
- [21] W. Hardyanto *et al.*, "Improving Sentiment Analysis with a Context-Aware RoBERTa-BiLSTM and Word2Vec Branch," *Scientific Journal of Informatics*, vol. 12, no. 4, pp. 785–796, Jan. 2026, doi: 10.15294/sji.v12i4.35918.
- [22] A. F. Wibowo and A. Susanto, "Analisis Sentimen Twitter Terhadap Isu Royalti Lagu di Industri Musik Indonesia Menggunakan Naive Bayes dan Support Vector Machine Berbasis TF-IDF," *Technology and Science (BITS)*, vol. 8, no. 1, 2026, doi: 10.47065/bits.v8i1.9842.
- [23] B. Bakirarar and A. H. Elhan, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research," *Turkiye Klinikleri Journal of Biostatistics*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.
- [24] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of SVM and Chi-Square Feature Selection for Sentiment Analysis of Indonesia's National Health Insurance Mobile Application," *Mathematics*, vol. 11, no. 17, Sep. 2023, doi: 10.3390/math11173765.
- [25] L. Ke, G. Hu, M. Zhao, Z. Liu, Z. Lv, and Y. Yang, "Brain tumor classification from MRI images using a multi-scale channel attention CNN integrated with SVM," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-36164-3.
- [26] C. Zhen *et al.*, "Grad-CAM based deep learning analytics for image-level colon disease classification based on graph neural networks and vision transformers," *Front. Physiol.*, vol. 17, 2026, doi: 10.3389/fphys.2026.1734299.
- [27] A. Althunibat, R. Amro, B. Hawashin, H. AlNuhait, S. Almanasra, and H. A. Al-Khawaja, "Automated Classification of User Requirements written in Arabic using machine learning algorithms," *Applied Mathematics and Information Sciences*, vol. 17, no. 6, pp. 1155–1170, 2023, doi: 10.18576/AMIS/170620.
- [28] Muhammad Mutawakkil Alallah, Mokhammad Amin Hariyadi, Triyo Supriatno, and E. Riana, "Implementasi BERT dan IndoRoBERTa untuk Klasifikasi Sentimen Opini Publik tentang Kecerdasan Buatan dalam Pendidikan di YouTube," *JURIKOM (Jurnal Riset Komputer)*, vol. 13, no. 2, pp. 666–678, Apr. 2026, doi: 10.30865/jurikom.v13i2.9697.