BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

Kinerja Metode Fine-Tuning IndoBERT untuk Klasifikasi Emosi Multi-Kelas pada Teks Informal Bahasa Indonesia

Haikal Fawwaz Karim[1,*], Adityo Permana Wibowo[2]

[1] Fakultas Sains dan Teknologi, Program Studi Informatika, Universitas Teknologi Yogyakarta, Sleman, Indonesia
[2] Fakultas Sains dan Teknologi, Program Studi Sistem Informasi, Universitas Teknologi Yogyakarta, Sleman, Indonesia
Email: [1,*]haikalfkarim@gmail.com, [2]adityopw@uty.ac.id
Email Penulis Korespondensi: haikalfkarim@gmail.com

Abstrak−Analisis emosi otomatis pada teks informal berbahasa Indonesia merupakan tugas yang menantang karena tingginya variasi linguistik, penggunaan bahasa gaul, dan singkatan. Penelitian ini berfokus pada pengembangan dan evaluasi model klasifikasi emosi yang akurat, yang dapat menjadi komponen dasar yang andal untuk berbagai aplikasi *Natural Language Processing* (NLP) yang relevan. Metode yang diusulkan adalah *fine-tuning* model bahasa *pre-trained* IndoBERT untuk mengklasifikasikan teks dari media sosial Twitter (X) ke dalam lima kelas emosi: 'anger' (marah), 'fear' (takut), 'happy' (senang), 'love' (cinta), dan 'sadness' (sedih). Sebuah dataset kustom yang terdiri dari 4.940 cuitan Twitter dibangun melalui proses *scraping* bertarget dan pelabelan yang tervalidasi secara statistik untuk memastikan relevansi dan keseimbangan data. Eksperimen menunjukkan bahwa setelah melalui tahapan pra-pemrosesan teks yang komprehensif, termasuk normalisasi menggunakan kamus singkatan kustom dan *stemming*, model yang di-*fine-tuning* mampu mencapai kinerja yang sangat tinggi. Hasil evaluasi pada data uji menunjukkan model berhasil mencapai akurasi sebesar 94% dan F1-score rata-rata tertimbang 0.94. Analisis kurva pembelajaran juga mengonfirmasi bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang baik. Hasil ini membuktikan bahwa pendekatan *fine-tuning* IndoBERT merupakan solusi yang sangat efektif dan andal untuk klasifikasi emosi pada domain teks informal bahasa Indonesia.

Kata Kunci: Klasifikasi Emosi; IndoBERT; Fine-Tuning; Teks Informal; Twitter

Abstract−Automatic emotion analysis on informal Indonesian texts is a challenging task due to high linguistic variation, the use of slang, and abbreviations. This research focuses on the development and evaluation of an accurate emotion classification model, which can serve as a core component various relevant Natural Language Processing (NLP) applications. The proposed method is the fine-tuning of the pre-trained language model IndoBERT to classify texts from the social media platform Twitter (X) into five emotion classes: anger, fear, happy, love, and sadness. A custom dataset consisting of 4,940 Twitter posts was built through a targeted scraping process and statistically validated labeling to ensure data relevance and balance. Experiments show that after undergoing a comprehensive text preprocessing stage, including normalization using a custom abbreviation dictionary and stemming, the fine-tuned model achieved very high performance. Evaluation results on the test data show the model successfully reached an accuracy of 94% and a weighted average F1-score of 0.94. Learning curve analysis also confirms that the model did not suffer from overfitting and possesses good generalization capabilities. These results demonstrate that the IndoBERT fine-tuning approach is a highly effective and reliable solution for emotion classification in the informal Indonesian text domain.

Keywords: Emotion Classification; IndoBERT; Fine-Tuning; Informal Text; Twitter

1. PENDAHULUAN

Era digital telah mentransformasi interaksi manusia, dengan media sosial seperti Twitter (X) menjadi ruang publik global tempat jutaan pengguna mencurahkan opini, berbagi pengalaman, dan mengungkapkan emosi[1]. Setiap hari, ratusan juta cuitan dihasilkan, menciptakan volume data tekstual tidak terstruktur yang sangat besar. Data ini, yang dikenal sebagai *User-Generated Content* (UGC), merupakan sumber daya yang begitu berharga untuk mengekstraksi wawasan mendalam mengenai sentimen publik, tren sosial, dan kondisi psikologis masyarakat[2]. Namun, mengekstraksi makna dari data ini bukanlah tugas yang mudah. Teks yang berasal dari media sosial memiliki karakteristik unik yang menjadi tantangan besar bagi pemrosesan bahasa alami/*Natural Language Processing* (NLP). Teks tersebut sangat informal, penuh dengan bahasa slang, singkatan non-standar (misalnya, "bgt", "yg", "gue"), kesalahan ketik, dan seringkali memiliki struktur kalimat yang tidak baku[3].

Secara historis, analisis teks otomatis sering berfokus pada analisis sentimen, yang mengklasifikasikan teks ke dalam polaritas sederhana seperti positif, negatif, atau netral[4]. Meskipun berguna, pendekatan ini seringkali tidak memadai untuk menangkap spektrum penuh dari perasaan manusia. Sebagai contoh, emosi 'anger' (marah) dan 'fear' (takut) keduanya memiliki polaritas negatif, tetapi memiliki implikasi yang benar-benar berbeda[5]. Untuk aplikasi yang lebih personal dan mendalam, seperti sistem rekomendasi yang adaptif, diperlukan pemahaman emosi yang lebih bernuansa[6]. Oleh karena itu, penelitian di bidang klasifikasi emosi multi-kelas (misalnya, membedakan antara senang, sedih, marah, takut, dan cinta) menjadi semakin krusial. Tantangan utamanya adalah bagaimana membangun sebuah model yang tidak hanya dapat memproses teks informal bahasa Indonesia, tetapi juga mampu memahami konteks kalimat secara menyeluruh untuk membedakan nuansa emosi yang perbedaannya acapkali tidak kentara.

Sebagai solusi atas tantangan pemahaman konteks, kemajuan dalam arsitektur *deep learning*, khususnya model *Transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*), telah menunjukkan kinerja yang luar biasa [7]. Berbeda dengan model sekuensial terdahulu (seperti RNN/*Recurrent Neural Network* atau LSTM/*Long-Short Term Memory*), *Transformer* mampu membaca seluruh kalimat sekaligus dan memahami hubungan kontekstual antara kata-kata melalui mekanisme *self-attention* [8]. Untuk domain bahasa Indonesia, IndoBERT hadir sebagai model bahasa *pre-trained* yang telah dilatih secara ekstensif pada korpus besar bahasa Indonesia (di antaranya adalah data berita,

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

Wikipedia, dan media sosial). IndoBERT menyediakan fondasi pemahaman bahasa yang kuat. Namun, model *pre-trained* ini bersifat umum dan perlu diadaptasi untuk tugas yang sangat spesifik. Solusi yang diusulkan dalam penelitian ini adalah menerapkan metode *fine-tuning* pada model IndoBERT. *Fine-tuning* adalah proses mengadaptasi pengetahuan umum dari model *pre-trained* untuk dispesialisasi dalam tugas tertentu yang dalam hal ini adalah klasifikasi emosi dengan melatihnya pada dataset yang relevan[9].

Penelitian terkait klasifikasi emosi dan sentimen di Indonesia telah berkembang pesat. Beberapa penelitian awal berfokus pada metode *machine learning* tradisional seperti Naïve Bayes dan *Support Vector Machine* (SVM) yang dikombinasikan dengan fitur leksikon [10], [11]. Meskipun menunjukkan hasil yang baik, metode ini sangat bergantung pada rekayasa fitur manual dan kamus sentimen yang terbatas. Penelitian yang lebih baru mulai mengadopsi *deep learning* karena terbukti lebih unggul dibandingkan dengan *machine learning*, terutama dalam mendeteksi emosi yang terkandung dalam teks yang ada di media sosial[12]. Misalnya, penelitian yang menggunakan *Long Short-Term Memory* (LSTM) dengan metode *word embedding* untuk klasifikasi sentimen pada *tweet* tentang Covid-19 Varian Omicron, menunjukkan peningkatan dalam menangkap urutan kata, tetapi masih terbatas pada polaritas positif-negatif[13]. Penelitian lain juga mencoba melakukan analisis sentimen dua kelas, yaitu positif dan negatif menggunakan *Convolutional Neural Network* (CNN), tetapi menghadapi kesulitan dalam menangani ambiguitas kata-kata dalam teks informal[14]. Di sisi lain, beberapa penelitian telah menerapkan model berbasis BERT. Sebuah penelitian membandingkan kinerja mBERT (BERT multilingual) dengan IndoBERT untuk mendeteksi hoaks berbahasa Indonesia, hasilnya menunjukkan bahwa IndoBERT memiliki keunggulan untuk klasifikasi teks spesifik dalam bahasa Indonesia dibandingkan dengan mBERT yang lebih umum karena bersifat multilingual[15]. Keunggulan IndoBERT juga telah divalidasi dalam tugas-tugas NLP bahasa Indonesia lainnya. Salah satu penelitian menggunakan IndoBERT untuk analisis sentimen yang terbukti lebih stabil dibandingkan dengan menggunakan CNN, tetapi dataset yang digunakan adalah berita formal, yang memiliki karakteristik sangat berbeda dari teks media sosial yang informal dan *noisy* (berisik)[16]. Penelitian lain berhasil mengimplementasikan IndoBERT untuk melakukan analisis sentimen multi-kelas pada opini publik tentang calon presiden di media sosial Twitter, meski terbatas pada kelas positif, negatif, dan netral[17].

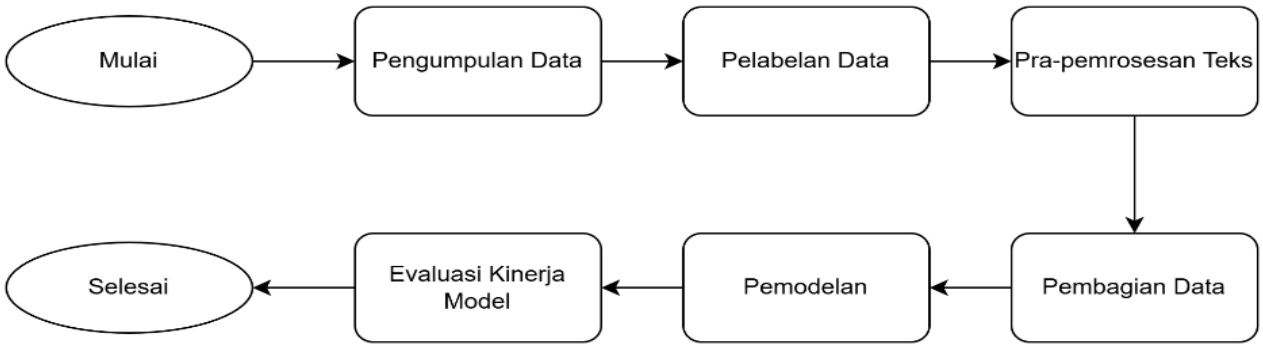
Dari tinjauan pustaka di atas, muncul sebuah *gap analysis* (celah penelitian) yang jelas: masih sedikit penelitian yang secara spesifik menguji dan mengevaluasi kinerja *fine-tuning* IndoBERT untuk tugas klasifikasi emosi multi-kelas (lima kelas) pada domain teks yang sangat informal dan tidak terstruktur seperti cuitan Twitter. Selain itu, banyak penelitian sebelumnya menggunakan dataset yang tidak seimbang (*imbalanced*), yang dapat menghasilkan model yang bias. Penelitian ini dirancang untuk mengisi celah tersebut dengan membangun dataset kustom yang seimbang dan menerapkan proses *fine-tuning* yang dioptimalkan.

Berdasarkan uraian di atas, tujuan dari penelitian ini adalah untuk (1) membangun sebuah dataset kustom berbahasa Indonesia dari Twitter untuk klasifikasi lima kelas emosi (‘anger’, ‘fear’, ‘happy’, ‘love’, ‘sadness’); (2) menerapkan dan mengoptimalkan metode *fine-tuning* pada model IndoBERT (indobenchmark/indobert-base-p1) melalui serangkaian tahapan pra-pemrosesan teks yang komprehensif; dan (3) menganalisis serta mengevaluasi kinerja model secara mendalam menggunakan metrik akurasi, presisi, recall, F1-score, dan *confusion matrix*. Harapan yang ingin dicapai adalah menghasilkan sebuah model klasifikasi emosi yang *robust* dan memiliki kinerja tinggi yang dapat dipertanggungjawabkan secara ilmiah. Keberhasilan model ini diharapkan dapat menjadi *benchmark* (tolok ukur) bagi penelitian sejenis dan dapat diimplementasikan sebagai komponen inti yang andal dalam berbagai sistem personalisasi cerdas atau aplikasi NLP lainnya.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimental untuk membangun dan mengevaluasi model klasifikasi emosi. Metodologi penelitian disusun dalam beberapa tahapan utama yang sistematis, meliputi: pengumpulan data, pelabelan data, pra-pemrosesan teks, pembagian data latih & data uji, pemodelan dengan *fine-tuning* IndoBERT, dan evaluasi kinerja model. Alur kerja dari penelitian ini diilustrasikan pada Gambar 1.



Gambar 1 Tahapan Penelitian

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

Dari Gambar 1 dapat dijelaskan:

a. Pengumpulan Data
Data mentah berupa 4.940 cuitan berbahasa Indonesia dikumpulkan dari platform media sosial Twitter (X) menggunakan teknik *scraping*. Untuk mengurangi potensi ketidakseimbangan kelas (*class imbalance*), proses *scraping* dilakukan secara bertarget untuk masing-masing dari lima kelas emosi: 'anger' (marah), 'fear' (takut), 'happy' (senang), 'love' (cinta), dan 'sadness' (sedih). Tiap-tiap emosi memiliki serangkaian kata kunci pencarian spesifik yang sering diasosiasikan dengan ekspresi emosi tersebut. Sebagai contoh, untuk emosi 'senang', kata kunci yang digunakan antara lain "seneng banget", "bahagia sekali", dan "bersyukur". Strategi ini bertujuan untuk mengurangi ketidakseimbangan kelas (*class imbalance*) sejak tahap awal pengumpulan data.

b. Pelabelan Data
Setelah data mentah terkumpul, proses pelabelan dilakukan untuk membuat *ground truth* bagi pelatihan model. Proses ini melibatkan dua orang pelabel (*rater*) yang memiliki pemahaman yang sama mengenai definisi operasional setiap kelas emosi, yaitu 'anger' (marah), 'fear' (takut), 'happy' (senang), 'love' (cinta), dan 'sadness' (sedih). Untuk memastikan objektivitas dan validitas *ground truth*, pelabelan dilakukan oleh 2 orang *rater* (pelabel). Konsistensi pelabelan diukur menggunakan metrik kesepakatan antar-pelabel (*Inter-Rater Agreement* - IRA). Dalam penelitian ini, sebuah studi validasi *post-hoc* dilakukan pada 400 sampel acak. Metrik yang digunakan adalah Cohen's Kappa (karena dua pelabel) dan didapatkan skor sebesar 0.70. Skor ini menunjukkan tingkat kesepakatan "Baik" (*Substantial*) antar pelabel, sehingga dataset yang dihasilkan dapat dianggap reliabel dan tidak bias secara subjektif [18], [19].

c. Pra-pemrosesan Teks
Teks dari media sosial bersifat sangat *noisy* (berisik) dan tidak terstruktur. Oleh karena itu, serangkaian tahapan pra-pemrosesan teks diperlukan untuk membersihkan dan mengubah teks dalam dataset menjadi format yang dapat dipahami oleh model. Tahapan yang dilakukan meliputi:
1. *Cleaning*: Menghapus elemen-elemen yang tidak relevan seperti URL, *mention* pengguna (@username), *hashtag* (#topic), serta karakter selain huruf dan spasi.
2. *Case Foldinbg*: Mengubah seluruh teks menjadi format huruf kecil (*lowercase*) untuk menyeragamkan kosakata.
3. *Normalization*: Melakukan normalisasi kata-kata tidak baku, seperti bahasa gaul dan singkatan, menjadi bentuk bakunya.
4. *Stopword Removal*: Menghapus kata-kata umum dalam Bahasa Indonesia yang tidak memiliki makna signifikan (misalnya, "yang", "di", "dan", "ini") menggunakan pustaka Sastrawi.
5. *Stemming*: Mengubah kata-kata berimbuhan menjadi bentuk kata dasarnya untuk mengurangi variasi kata dengan makna inti yang sama (misalnya, "menulis", "ditulis", "tulisan" diubah menjadi "tulis"). Proses ini juga menggunakan pustaka Sastrawi.

d. Pembagian Data
Pembagian data dalam penelitian ini dilakukan dengan membagi dataset menjadi dua bagian: 80% data latih dan 20% data uji. Pembagian ini dilakukan menggunakan metode *stratified sampling* untuk memastikan proporsi setiap kelas emosi pada data latih dan data uji tetap identik dengan distribusi pada data keseluruhan. Dalam implementasi penelitian ini, 20% data uji (988 sampel) menjalankan peran ganda. Peran pertamanya adalah sebagai data validasi (*evaluation dataset*) selama proses pelatihan. Data ini digunakan untuk memantau kinerja model (eval_loss, eval_f1) di akhir setiap *epoch*, yang menginformasikan mekanisme EarlyStopping dan load_best_model_at_end. Kami mengakui bahwa dalam metodologi yang ketat, set validasi (yang digunakan untuk *early stopping* dan pemilihan model) idealnya terpisah dari set uji final. Namun, mengingat ukuran dataset kustom yang relatif terbatas (4.940 sampel), pendekatan 80/20 ini dipilih untuk memaksimalkan jumlah data latih, sekaligus menyediakan set evaluasi yang cukup stabil (988 sampel). Oleh karena itu, laporan kinerja final (di Bab 3) disajikan pada set 20% ini, dengan pemahaman bahwa set tersebut juga telah digunakan untuk pemilihan model.

e. Pemodelan
Penelitian ini menggunakan arsitektur *Transformer* melalui model *pre-trained* IndoBERT yang merupakan varian dari BERT.
1. Latar Belakang Arsitektur: BERT dan IndoBERT
BERT merupakan singkatan dari *Bidirectional Encoder Representations from Transformers*. BERT adalah sebuah model representasi bahasa, yakni model yang dirancang untuk menghasilkan representasi kata atau kalimat yang memahami konteks[20]. Berbeda dengan model-model lain yang memproses teks secara satu arah (kiri-ke-kanan atau kanan-ke-kiri), BERT menawarkan inovasi untuk melakukan pra-pelatihan representasi mendalam yang bersifat dua arah (*bidirectional*)[21]. Hal ini dicapai dengan mengondisikan model pada konteks kiri dan kanan secara bersamaan pada semua lapisannya. Ini artinya, BERT mampu memahami makna sebuah kata dengan melihat keseluruhan kalimat, yang memberinya pemahaman kontekstual yang jauh lebih unggul. Meskipun model BERT asli sangat kuat, ia dilatih secara dominan pada teks berbahasa Inggris. Hal ini menjadi tantangan tatkala ia diimplementasikan pada bahasa lain yang memiliki sintaksis yang berbeda, seperti bahasa Indonesia seperti pada penelitian ini. Apalagi jika ditambah dengan keragaman linguistik di Indonesia yang begitu beragam dengan lebih dari 700 bahasa dan dialek, sehingga dapat memunculkan adanya praktik *code-switching* (campur kode bahasa) yang sangat umum. Model bahasa yang tidak dilatih secara spesifik pada data lokal akan kesulitan menangani kerumitan ini[22]. Untuk mengatasi tantangan yang dihadapi oleh BERT konvensional, dikembangkanlah IndoBERT. IndoBERT adalah implementasi dari arsitektur BERT yang secara khusus melalui proses pra-pelatihan

dari awal menggunakan korpus data raksasa bahasa Indonesia yang berisi lebih dari 4 miliar kata yang berasal dari berbagai sumber teks bahasa Indonesia, termasuk artikel berita, media sosial, blog, dan Wikipedia. Dengan dilatih pada miliaran kata yang relevan dengan konteks lokal, IndoBERT memiliki pemahaman yang jauh lebih mendalam mengenai tata bahasa, kosakata, idiom, dan bahkan bahasa gaul yang umum digunakan di Indonesia [23], [24].

2. Model Dasar dan Proses *Fine-Tuning*
   Model dasar yang digunakan dalam penelitian ini adalah indobenchmark/indobert-base-p1. Kemampuannya untuk memahami konteks kalimat dalam Bahasa Indonesia menjadikannya fondasi yang sangat kuat untuk tugas-tugas NLP. Proses *fine-tuning* kemudian diterapkan, yang merupakan tahap di mana model IndoBERT diadaptasi untuk tugas spesifik klasifikasi emosi. Bobot (*weights*) dari model *pre-trained* tidak dimulai dari nol, melainkan disesuaikan kembali menggunakan dataset emosi yang telah kita siapkan. Lapisan klasifikasi baru ditambahkan di atas model IndoBERT untuk memetakan representasi teks ke salah satu dari lima kelas emosi. Dataset dibagi menjadi 80% data latih dan 20% data uji.

Untuk meningkatkan kemampuan generalisasi dan mencegah *overfitting*, *hyperparameter model* dioptimalkan. Model indobenchmark/indobert-base-p1 dilatih menggunakan optimizer AdamW (*Adaptive Moment Estimation with Weight Decay)* dengan *learning rate* yang diatur ke 5e-06. Secara spesifik, tingkat *dropout* pada *attention* dan *hidden layer* ditingkatkan menjadi 0.3 untuk regularisasi. Proses pelatihan dijalankan dengan *batch size* 16 dan direncanakan untuk maksimal 5 epoch. Untuk memastikan model *tidak overfitting* dan mendapatkan hasil terbaik, diterapkan dua kebijakan: *Early Stopping* dengan patience 2 (berhenti jika tidak ada perbaikan F1-score setelah 2 *epoch*) dan load_best_model_at_end yang akan memuat kembali bobot model dengan metrik F1-score validasi tertinggi.

f. Evaluasi Kinerja Model
   Kinerja model dievaluasi menggunakan empat metrik standar untuk tugas klasifikasi. Metrik ini didasarkan pada empat hasil dasar dari *Confusion Matrix* untuk setiap kelas: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN)[25].
   Metrik yang pertama adalah akurasi (*accuracy*). Akurasi mengukur proporsi prediksi yang benar secara keseluruhan, dihitung menggunakan rumus (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Namun, untuk dataset yang mungkin sedikit tidak seimbang, Akurasi saja tidak cukup. Oleh karena itu, digunakan Presisi, *Recall*, dan *F1-Score* yang dihitung per kelas.
Presisi (*Precision*) mengukur seberapa akurat prediksi positif yang dibuat oleh model. Dihitung menggunakan rumus (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall (disebut juga sensitivitas) mengukur seberapa baik model menemukan semua sampel positif yang relevan. Dihitung menggunakan rumus (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

*F1-Scor*e adalah rata-rata harmonis dari Presisi dan *Recall*, yang memberikan satu metrik tunggal yang menyeimbangkan kedua nilai tersebut. Ini menjadi metrik utama untuk mengevaluasi performa model pada setiap kelas emosi. Dihitung menggunakan rumus (4):

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pengumpulan data dilakukan dengan teknik *scraping* menggunakan bahasa pemrograman python dan pustaka Selenium. Data yang telah terkumpul berjumlah 4.940 cuitan yang contohnya dapat dilihat pada Tabel 1.

Tabel 1. Contoh Data Hasil *Scraping*

No	Tweet
1	Seneng banget pasti ini kak
2	seneng banget dipercaya cust
3	Manusia bijak adalah saat sukses bisa bersyukur saat gagalpun tetap bersyukur. Karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur
4	Dari gesturenya kelihatan Mbak Glory dam Mas Brein (yang kabarnya dapat kesempatan menginap dan berenang di Aman NY), bahagia sekali, dapat bertemu dengan Presiden Prabowo Subianto.dan rombongan.
5	guee bersyukur sekali ga di jam 2 ini match biasanyaa selalu ketiduran

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

Tabel 1 menyajikan lima contoh data mentah (*raw data*) yang berhasil dikumpulkan. Penjelasan ini mengilustrasikan karakteristik teks Twitter yang informal dan *noisy*, seperti penggunaan bahasa gaul ("seneng banget", "guee") dan tanda baca yang tidak standar, yang menjadi tantangan utama dalam penelitian ini.

3.2 Pelabelan Data

Data yang telah dikumpulkan pada tahap sebelumnya masih berupa teks mentah tanpa label. Untuk membuat dataset yang siap diproses, dilakukan pelabelan data. Hasil dari tahapan ini adalah dataset final yang digunakan untuk semua proses selanjutnya. Kelas yang digunakan adalah ‘anger’, ‘fear’, ‘happy’, ‘love’, dan ‘sadness’. Contoh hasil pelabelan dapat dilihat pada Tabel 2.

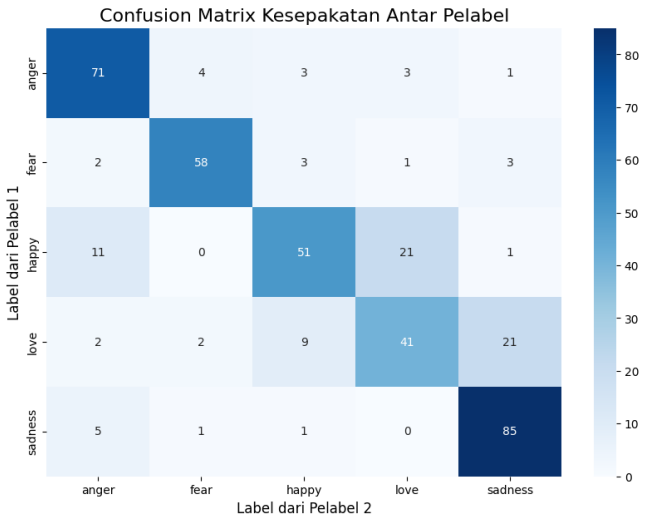
Tabel 2. Contoh Data Setelah Pelabelan

No	Label	Tweet
1	happy	Seneng banget pasti ini kak
2	happy	seneng banget dipercaya cust
3	happy	Manusia bijak adalah saat sukses bisa bersyukur saat gagalpun tetap bersyukur. Karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur
4	happy	Dari gesturenya kelihatan Mbak Glory dam Mas Brein (yang kabarnya dapat kesempatan menginap dan berenang di Aman NY), bahagia sekali, dapat bertemu dengan Presiden Prabowo Subianto.dan rombongan.
5	happy	guee bersyukur sekali ga di jam 2 ini match biasanyaa selalu ketiduran

Tabel 2 memberikan contoh konkrit dari hasil proses pelabelan data. Tabel ini secara rinci menunjukkan bagaimana setiap data teks (pada kolom 'Tweet') yang sebelumnya mentah, kini telah dipasangkan dengan salah satu dari lima kelas emosi yang ditentukan (pada kolom 'Label'). Sebagai contoh, kelima sampel yang ditampilkan di tabel ini diidentifikasi memiliki emosi 'happy' (seperti pada teks "Seneng banget pasti ini kak" dan "guee bersyukur sekali"). Pasangan data (teks, label) inilah yang membentuk *ground truth* (kebenaran dasar) final untuk 4.940 sampel, yang validitasnya akan dibuktikan secara statistik pada bagian selanjutnya (studi IRA) sebelum digunakan untuk pra-pemrosesan.

Sebelum dataset penuh digunakan untuk pelatihan, hasil dari studi validasi *Inter-Rater Agreement* (IRA) disajikan. Seperti yang dijelaskan dalam metodologi penelitian, studi ini dilakukan pada 400 sampel acak menggunakan dua pelabel untuk menghitung Cohen's Kappa. Hasil perhitungan menunjukkan skor Kappa sebesar 0.70, yang mengindikasikan tingkat kesepakatan 'Baik' (*Substantial*). Skor ini mengonfirmasi bahwa *ground truth* asli (label orisinal) yang digunakan untuk melatih model memiliki reliabilitas dan objektivitas yang tinggi, sehingga layak digunakan untuk pelatihan dan evaluasi.

Untuk menganalisis lebih lanjut pola kesepakatan antar pelabel, *confusion matrix* dari studi IRA disajikan pada Gambar 2.

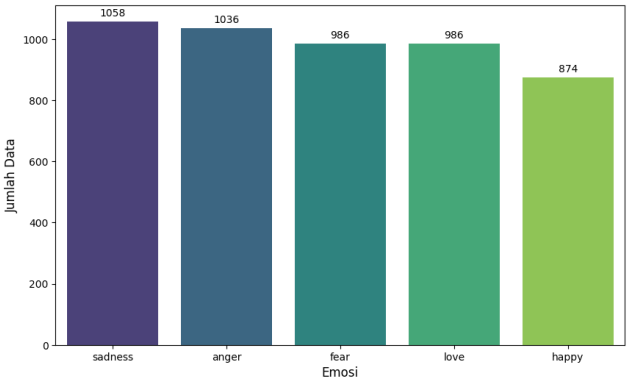


Gambar 2 *Confusion Matrix* Kesepakatan Antar Pelabel (IRA)

Dari Gambar 2, dapat dilihat bahwa kesepakatan tertinggi terjadi pada kelas ‘sadness’ (85 dari 92 sepakat) dan ‘anger’ (71 dari 82 sepakat). Namun, terdapat pola ketidaksepakatan yang menarik dan signifikan secara statistik. Pertama, terjadi kebingungan antara kelas ‘happy’ dan ‘love’: terdapat 21 kasus di mana Pelabel 1 memberi label ’happy’, tetapi Pelabel 2 memberi label ’love’. Kedua, terdapat ambiguitas antara ‘love’ dan ‘sadness’: ada 21 kasus di mana Pelabel 1 memberi label ‘love’, tetapi Pelabel 2 memberi label ‘sadness’. Ketiga, terdapat pula 11 kasus di mana Pelabel 1 menilai ‘happy’, sedangkan Pelabel 2 menilai ‘anger’. Pola ketidaksepakatan antar-manusia ini sangat penting, karena

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

menunjukkan adanya ambiguitas alami (*natural ambiguity*) dalam ekspresi emosi tersebut pada teks informal, yang kemungkinan besar juga akan menjadi tantangan bagi model IndoBERT.
Distribusi kelas dari dataset final disajikan pada Gambar 3.



Gambar 3. Distribusi Kelas Pada Dataset

Seperti yang terlihat pada Gambar 3, kelas 'sadness' memiliki jumlah sampel terbanyak (1.058 sampel), diikuti oleh 'anger' (1.036), 'fear' (986), 'love' (986), dan 'happy' (874). Meskipun strategi *scraping* bertarget telah diterapkan, ketidakseimbangan minor ini wajar terjadi karena perbedaan frekuensi penggunaan kata kunci pencarian di platform. Ketidakseimbangan ini tidak ekstrem dan dapat dikelola. Justru, hal ini memvalidasi pentingnya penggunaan metode *stratify* pada saat pembagian data latih dan uji, untuk memastikan bahwa proporsi minoritas (seperti kelas 'happy') tetap terwakili secara adil di kedua dataset, sehingga evaluasi model tetap objektif.

3.3 Pra-pemrosesan Teks

3.3.1 *Cleaning*

Tahap pertama dalam pra-pemrosesan adalah *cleaning*, tugasnya untuk membersihkan teks dari tanda baca, emoji, dan sebagainya. Hasil pembersihan (*cleaning*) dapat dilihat pada Tabel 3.

Tabel 3. Contoh Data Setelah *Cleaning*

No	Teks Asli	Teks *Cleaning*
1	Seneng banget pasti ini kak	Seneng banget pasti ini kak
2	seneng banget dipercaya cust	seneng banget dipercaya cust
3	Manusia bijak adalah saat sukses bisa bersyukur saat gagalpun tetap bersyukur. Karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur	Manusia bijak adalah saat sukses bisa bersyukur saat gagalpun tetap bersyukur Karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur
4	Dari gesturenya kelihatan Mbak Glory dam Mas Brein (yang kabarnya dapat kesempatan menginap dan berenang di Aman NY), bahagia sekali, dapat bertemu dengan Presiden Prabowo Subianto.dan rombongan.	Dari gesturenya kelihatan Mbak Glory dam Mas Brein yang kabarnya dapat kesempatan menginap dan berenang di Aman NY bahagia sekali dapat bertemu dengan Presiden Prabowo Subiantodan rombongan
5	guee bersyukur sekali ga di jam 2 ini match biasanyaa selalu ketiduran	guee bersyukur sekali ga di jam ini match biasanyaa selalu ketiduran

Tabel 3 menyajikan perbandingan antara teks asli dengan hasil dari tahapan *Cleaning*. Seperti yang dapat dilihat, proses ini berhasil menghapus elemen-elemen noise yang tidak relevan. Contohnya pada baris ke-4, tanda baca seperti tanda kutip (") dan koma (,) serta tanda kurung ((,)) dan titik (.) telah dihilangkan, menghasilkan teks yang lebih bersih dan siap untuk tahapan *Case Folding*.

3.3.2 *Case Folding*

Tahap lanjutan setelah dilakukannya *cleaning* adalah *case folding*, ini berguna untuk membuat teks menjadi seragam, dari sebelumnya mengandung huruf kapital distandarkan seluruhnya menjadi huruf kecil. Contoh data yang diproses dengan *case folding* dapat dilihat pada Tabel 4.

Tabel 4. Contoh Data Setelah *Case Folding*

No	Teks *Cleaning*	Teks *Case Folding*
1	Seneng banget pasti ini kak	seneng banget pasti ini kak
2	seneng banget dipercaya cust	seneng banget dipercaya cust
3	Manusia bijak adalah saat sukses bisa bersyukur	manusia bijak adalah saat sukses bisa bersyukur saat

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

No	Teks *Cleaning*	Teks *Case Folding*
	saat gagalpun tetap bersyukur Karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur	gagalpun tetap bersyukur karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur
4	Dari gesturenya kelihatan Mbak Glory dam Mas Brein yang kabarnya dapat kesempatan menginap dan berenang di Aman NY bahagia sekali dapat bertemu dengan Presiden Prabowo Subiantodan rombongan	dari gesturenya kelihatan mbak glory dam mas brein yang kabarnya dapat kesempatan menginap dan berenang di aman ny bahagia sekali dapat bertemu dengan presiden prabowo subiantodan rombongan
5	guee bersyukur sekali ga di jam ini match biasanyaa selalu ketiduran	guee bersyukur sekali ga di jam ini match biasanyaa selalu ketiduran

Tabel 4 mendemonstrasikan hasil dari tahap *Case Folding*. Tujuan dari langkah ini adalah untuk menyeragamkan seluruh teks ke dalam format huruf kecil (*lowercase*), menghilangkan ambiguitas yang mungkin timbul karena perbedaan penggunaan huruf kapital. Perubahan paling jelas dapat diamati pada baris ke-3 dan ke-4. Pada kolom 'Teks Cleaning' baris ke-3, kata "Manusia", "Karena", "Kekayaan", dan "Kebahagian" diawali dengan huruf kapital. Demikian pula pada baris ke-4, nama dan tempat seperti "Mbak Glory", "Mas Brein", "Aman NY", dan "Presiden Prabowo" juga menggunakan huruf kapital. Setelah proses *Case Folding* (di kolom 'Teks *Case Folding*'), semua kata tersebut telah berhasil diubah menjadi huruf kecil (misalnya, "manusia", "karena", "mbak glory", "presiden prabowo"). Penyeragaman ini sangat penting sebelum melangkah ke tahap *Normalisasi* dan *Stemming*.

3.3.3 *Normalization*

Setelah case *folding* telah berhasil dilakukan, selanjutnya dilakukan *normalization* (normalisasi). Normalisasi kata merupakan hal yang sangat penting untuk mengatasi tingginya variasi bahasa gaul dan singkatan dalam data Twitter. Proses normalisasi dilakukan dengan memetakan kata-kata tidak baku ke bentuk bakunya. Untuk melakukan ini, sebuah kamus singkatan kustom yang disimpan dalam berkas berekstensi .csv (kamus_singkatan.csv) dibangun dan digunakan. Kamus ini berisi lebih dari 2.000 entri yang dikumpulkan dari analisis dataset dan sumber bahasa gaul Indonesia, yang mencakup singkatan umum (misalnya, "bgt" diubah menjadi "sangat", "yg" menjadi "yang") dan kata ganti informal (misalnya, "gue" diubah menjadi "saya"). Contoh data hasil *normalization* dapat dilihat pada Tabel 5.

Tabel 5 Contoh Data Setelah *Normalization*

No	Teks *Case Folding*	Teks *Normalization*
1	seneng banget pasti ini kak	senang sangat pasti ini kakak
2	seneng banget dipercaya cust	senang sangat dipercaya cust
3	manusia bijak adalah saat sukses bisa bersyukur saat gagalpun tetap bersyukur karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur	manusia bijaksana adalah saat sukses dapat bersyukur saat gagalpun tetap bersyukur karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur
4	dari gesturenya kelihatan mbak glory dam mas brein yang kabarnya dapat kesempatan menginap dan berenang di aman ny bahagia sekali dapat bertemu dengan presiden prabowo subiantodan rombongan	dari gesturenya kelihatan kakak glory dam kakak brein yang kabarnya dapat kesempatan menginap dan berenang di aman ny bahagia sekali dapat bertemu dengan presiden prabowo subiantodan rombongan
5	guee bersyukur sekali ga di jam ini match biasanyaa selalu ketiduran	saya bersyukur sekali tidak di jam ini match biasanyaa selalu tertidur

Efektivitas tahap normalisasi, yang hasilnya disajikan pada Tabel 5, dapat terlihat jelas dengan merujuk pada beberapa contoh. Dengan menggunakan kamus kustom yang telah disiapkan, kata-kata informal yang umum seperti 'seneng' dan 'banget' (terlihat di baris ke-1 dan ke-2) berhasil distandarkan menjadi 'senang' dan 'sangat'. Contoh yang lebih signifikan terlihat pada baris ke-5, di mana kata ganti informal 'guee' diubah menjadi 'saya' dan singkatan 'ga' diubah menjadi 'tidak'. Selain itu, proses ini juga memperbaiki ejaan yang ambigu seperti pada baris ke-3, di mana 'bijak' diubah menjadi 'bijaksana'. Dengan memetakan ribuan variasi bahasa gaul dan singkatan ini ke bentuk baku yang konsisten, dataset menjadi jauh lebih seragam dan bersih, sehingga siap untuk tahap *stopword removal*.

3.3.4 *Stopword Removal*

Setelah teks berhasil dinormalisasi, maka tahap selanjutnya dalam penelitian ini adalah *stopword removal*, yaitu tahap untuk membuang kata-kata yang tidak terlalu diperlukan, biasanya mencakup konjungsi, seperti “dan”, “tetapi”, “atau”, kemudian juga preposisi, seperti “ini”, “itu”, kata ganti seperti “saya”, “kamu”, “dia”, dan sebagainya yang tidak memberikan makna penting dalam kalimat. Contoh data hasil *stopword removal* dapat dilihat pada Tabel 6.

Tabel 6 Contoh Data Setelah *Stopword Removal*

No	Teks *Normalization*	Teks *Stopword Removal*
1	senang sangat pasti ini kakak	senang sangat ini kakak
2	senang sangat dipercaya cust	senang sangat dipercaya cust

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

No	Teks *Normalization*	Teks *Stopword Removal*
3	manusia bijaksana adalah saat sukses dapat bersyukur saat gagalpun tetap bersyukur karena sesungguhnya kekayaan dan kebahagian sejati ada didalam rasa syukur	manusia bijaksana sukses bersyukur saat gagalpun tetap bersyukur sesungguhnya kekayaan kebahagian sejati didalam rasa syukur
4	dari gesturenya kelihatan kakak glory dam kakak brein yang kabarnya dapat kesempatan menginap dan berenang di aman ny bahagia sekali dapat bertemu dengan presiden prabowo subiantodan rombongan	gesturenya kelihatan kakak glory dam kakak brein kabarnya kesempatan menginap berenang aman ny bahagia sekali bertemu presiden prabowo subiantodan rombongan
5	saya bersyukur sekali tidak di jam ini match biasanyaa selalu tertidur	bersyukur sekali di jam match biasanyaa selalu tertidur

Tahap *Stopword Removal*, yang hasilnya ditunjukkan pada Tabel 6, bertujuan untuk mengurangi *noise* lebih lanjut dengan menghapus kata-kata umum (*stopword*) yang tidak memiliki makna emosional signifikan. Dapat dilihat pada baris ke-1 bahwa preposisi 'ini' dihilangkan. Contoh yang lebih jelas ada pada baris ke-3, di mana kata-kata seperti 'adalah', 'saat', 'dapat', 'karena', 'dan', 'ada', 'didalam' dihilangkan, sehingga kalimat menjadi lebih padat ("manusia bijaksana sukses bersyukur gagalpun tetap bersyukur sesungguhnya kekayaan kebahagian sejati rasa syukur"). Demikian pula, pada baris ke-5, kata ganti 'saya' dan preposisi 'tidak' dan 'di' telah dihapus. Dengan menghilangkan kata-kata non-informatif ini, teks yang tersisa menjadi lebih fokus pada kata-kata inti yang mengandung sentimen, yang sangat penting untuk tahap akhir, yaitu *Stemming*.

3.3.5 *Stemming*

Tahapan *stemming* adalah tahapan pra-pemrosesan teks terakhir dalam penelitian ini. *Stemming* bertindak untuk mengembalikan setiap kata dikembalikan ke dalam bentuk kata dasarnya. Misalnya, "bersyukur" menjadi "syukur", yang membantu model mengonsolidasikan kata-kata dengan makna inti yang sama. Contoh hasil data setelah *stemming* ada pada Tabel 7.

Tabel 7 Contoh Data Setelah *Stemming*

No	Teks *Stopword Removal*	Teks *Stemming*
1	senang sangat ini kakak	senang sangat ini kakak
2	senang sangat dipercaya cust	senang sangat percaya cust
3	manusia bijaksana sukses bersyukur saat gagalpun tetap bersyukur sesungguhnya kekayaan kebahagian sejati didalam rasa syukur	manusia bijaksana sukses syukur saat gagal tetap syukur sungguh kaya kebahagian sejati dalam rasa syukur
4	gesturenya kelihatan kakak glory dam kakak brein kabarnya kesempatan menginap berenang aman ny bahagia sekali bertemu presiden prabowo subiantodan rombongan	gesturenya lihat kakak glory dam kakak brein kabar sempat inap renang aman ny bahagia sekali temu presiden prabowo subiantodan rombong
5	bersyukur sekali di jam match biasanyaa selalu tertidur	syukur sekali di jam match biasanyaa selalu tidur

Tabel 7 mengilustrasikan langkah terakhir dan paling transformatif dalam *pipeline* pra-pemrosesan, yaitu *Stemming*. Tujuan dari tahap ini adalah untuk mengonsolidasikan berbagai kata berimbuhan ke dalam satu kata dasar (kata akar) yang seragam. Efek dari proses ini sangat terlihat di seluruh sampel. Pada baris ke-3, kata "bersyukur" dan "sesungguhnya" berhasil diubah menjadi kata dasarnya, "syukur" dan "sungguh". Contoh yang lebih drastis terlihat pada baris ke-4, di mana kata-kata kerja seperti "kelihatan", "menginap", "berenang", dan "bertemu" semuanya berhasil dikembalikan ke akarnya menjadi "lihat", "inap", "renang", dan "temu". Demikian pula pada baris ke-5, "bersyukur" menjadi "syukur" dan "tertidur" menjadi "tidur". Hasil dari kolom 'Teks Stemming' ini adalah *corpus* akhir yang bersih dan terstandarisasi, yang siap digunakan untuk melatih model IndoBERT.

Untuk memvalidasi lebih lanjut kualitas dataset yang telah dibersihkan, dilakukan analisis data eksploratif menggunakan *word cloud* untuk setiap kelas emosi, seperti yang ditunjukkan pada Gambar 3.

Gambar 4. Visualisasi *Wordcloud* Setiap Kelas Emosi

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

Hasil visualisasi pada Gambar 4 memberikan beberapa pengetahuan yang penting. Pertama, keberhasilan *pipeline* pra-pemrosesan terkonfirmasi secara visual; kata-kata yang dominan adalah kata-kata inti yang penuh dengan makna emosional, bukan *stopword*. Kedua, validitas dataset terbukti, di mana setiap kelas emosi secara jelas didominasi oleh kosakata yang relevan. Kelas 'anger' didominasi oleh kata seperti "sial" dan "muak". Kelas 'fear' didominasi oleh "takut", "ngeri", dan "merinding". Kelas 'happy' oleh "senang" dan "syukur". Kelas 'love' oleh "cinta", "sayang", dan "haru". Dan kelas 'sadness' oleh "sedih" dan "kecewa". Visualisasi ini menegaskan bahwa data yang telah disiapkan memiliki makna emosional yang kuat dan siap untuk digunakan dalam pelatihan model.

3.4 Pembagian Data

Dataset yang telah bersih (total 4.940 sampel) dibagi menjadi dua bagian: 80% data latih (3.952 sampel) dan 20% data uji (988 sampel). Data latih dijadikan sebagai bahan pembelajaran bagi model IndoBERT, sedangkan data uji adalah data untuk mengetes kinerja model IndoBERT. Distribusi kelas di dalam data latih dan data uji ditunjukkan dalam Tabel 8.

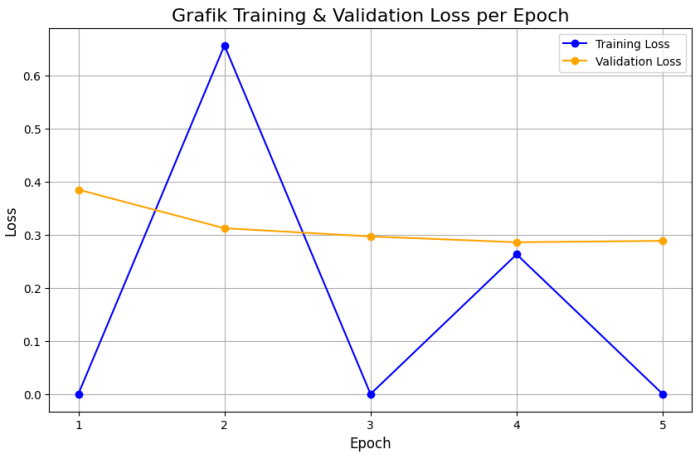
Tabel 8 Distribusi Sampel Data Latih dan Data Uji per Kelas

No	Kelas Emosi	Data Latih (80%)	Data Uji (20%)	Total
1	anger	829	207	1036
2	fear	789	197	986
3	happy	699	175	874
4	love	789	197	986
5	sadness	846	212	1058
Total		3952	988	4940

Tabel 8 menyajikan hasil akhir dari proses pembagian data (80% latih dan 20% uji) yang telah dijelaskan pada paragraf sebelumnya. Tabel ini secara kuantitatif membuktikan bahwa metode *stratified sampling* yang digunakan telah berhasil. Dapat diamati bahwa proporsi dari setiap kelas emosi dipertahankan secara ketat di kedua dataset. Sebagai contoh, kelas happy, yang merupakan kelas minoritas (17.7% dari total data), secara proporsional terbagi menjadi 699 sampel latih (17.7% dari data latih) dan 175 sampel uji (17.7% dari data uji). Hal yang sama berlaku untuk kelas mayoritas seperti sadness (21.4% dari total), yang juga terbagi secara proporsional (846 latih dan 212 uji). Distribusi yang seimbang ini sangat krusial untuk memastikan bahwa model dilatih pada *dataset* yang representatif dan dievaluasi secara adil pada data uji yang memiliki proporsi kelas yang identik.

3.5 Pemodelan

Stabilitas proses pelatihan model dievaluasi melalui kurva *Training & Validation Loss* per *epoch*, seperti yang ditunjukkan pada Gambar 4. Grafik ini memplot nilai kesalahan model pada data latih (*Training Loss*) dan data validasi (*Validation Loss*) di setiap akhir *epoch*.



Gambar 5 Grafik *Training* dan *Validation Loss*

Seperti yang terlihat pada Gambar 5, kedua kurva, baik *Training Loss* (biru) maupun *Validation Loss* (oranye), sama-sama menunjukkan tren menurun yang konsisten. Yang terpenting, kedua kurva bergerak sangat berdekatan dan tidak menunjukkan adanya celah (*gap*) yang melebar. Hal ini merupakan indikator kuat bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang sangat baik. Proses pelatihan yang stabil ini merupakan hasil dari penerapan *hyperparameter* yang tepat, termasuk peningkatan *dropout rate* menjadi 0.3 dan penggunaan *learning rate* yang lebih rendah (5e-6), yang berhasil mencegah model terlalu menghafal data latih.

3.6 Evaluasi Kinerja Model

Proses *fine*-tuning model IndoBERT pada dataset klasifikasi emosi yang telah disiapkan menunjukkan hasil yang sangat positif. Model dilatih selama 5 *epoch* dan berhasil mencapai kinerja yang sangat tinggi pada data uji yang belum pernah

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 63-74
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.850

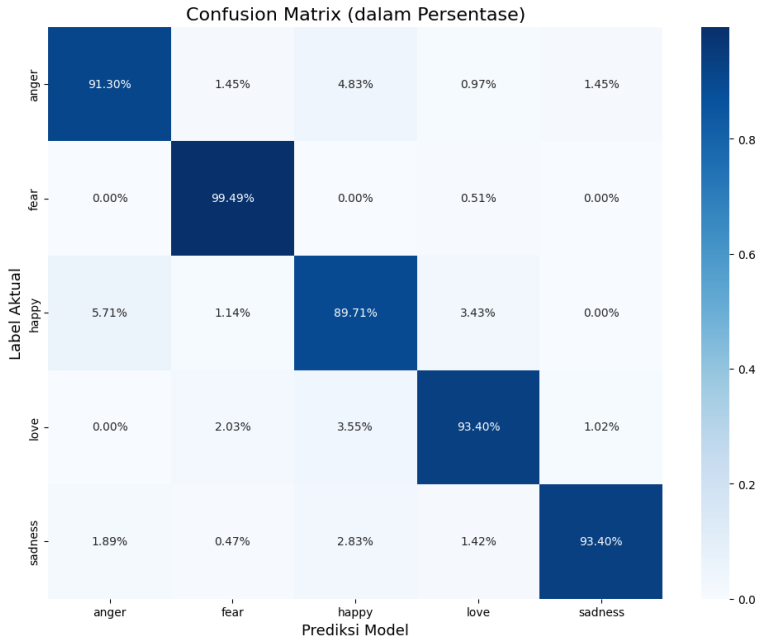
dilihat sebelumnya. Evaluasi akhir dilakukan pada 988 sampel data uji. Hasil evaluasi disajikan secara rinci pada Tabel 9. Secara keseluruhan, model yang di-*fine-tuning* berhasil mencapai akurasi sebesar 94%. Nilai *F1-Score* rata-rata tertimbang (*weighted avg*) juga mencapai 0.94, yang mengindikasikan bahwa model memiliki performa yang sangat baik dan seimbang di semua kelas emosi.

Tabel 9. Hasil Evaluasi Final pada Data Uji

Label	Precision	Recall	F1-Score	Support
anger	0.93	0.91	0.92	207
fear	0.95	0.99	0.97	197
happy	0.87	0.9	0.88	175
love	0.94	0.93	0.94	197
sadness	0.98	0.93	0.95	212
accuracy			0.94	988
macro avg	0.93	0.93	0.93	988
weighted avg	0.94	0.94	0.94	988

Tabel 9 menunjukkan bahwa performa tertinggi dicapai pada kelas ‘fear’ (F1-Score 0.97), di mana model mampu mengidentifikasi hampir semua sampel dengan benar. Performa terendah, meskipun masih tergolong sangat baik, tercatat pada kelas happy (*F1-Score* 0.88). Hal ini kemungkinan menunjukkan bahwa ekspresi kebahagiaan dalam teks informal memiliki variasi linguistik yang lebih luas atau lebih sering bersifat ambigu dibandingkan emosi lainnya.

Untuk memahami pola kesalahan yang dibuat oleh model, *Confusion Matrix* disajikan pada Gambar 5. Matriks ini memvisualisasikan perbandingan antara label aktual dengan label yang diprediksi oleh model dalam bentuk persentase.



Gambar 6 *Confusion Matrix* Penelitian

Garis diagonal *confusion matrix* pada Gambar 6 menunjukkan persentase prediksi yang benar, yang nilainya sangat dominan (berkisar antara 89.71% hingga 99.49%). Angka di luar diagonal menunjukkan kesalahan klasifikasi. Untuk label ‘anger’, 4,83% sampel salah diprediksi sebagai ‘happy’, 1,45% sebagai ‘fear’, 1,45% sebagai ‘sadness’, dan 0,97% sebagai ‘love’; untuk ‘fear’ hampir semua benar, hanya 0,51% yang tersasar ke ‘love’; untuk ‘happy’, kesalahan terbesar adalah 5,71% dikira ‘anger’, disusul 3,43% ke ‘love’ dan 1,14% ke ‘fear’; untuk ‘love’, 3,55% tertukar menjadi ‘happy’, 2,03% menjadi ‘fear’ dan 1,02% menjadi ‘sadness’; sedangkan untuk ‘sadness’, 2,83% salah ke ‘happy’, 1,89% ke ‘anger’, 1,42% ke ‘love’ dan 0,47% ke ‘fear’. Pola ini menunjukkan bahwa kebingungan paling signifikan terjadi antara pasangan ‘happy’ dengan ‘anger’ dan ‘happy’ dengan ‘love’, sedangkan untuk ‘fear’ nyaris tidak tertukar dengan kelas lain. Menariknya, pola kebingungan model ini (di mana 'happy' sering tertukar dengan 'anger' dan 'love') mencerminkan pola ambiguitas yang sama dengan yang ditemukan antar pelabel manusia pada Gambar 2, di mana ketidaksepakatan juga banyak terjadi antara kelas-kelas tersebut.

4. KESIMPULAN

Penelitian ini berhasil menunjukkan efektivitas metode *fine-tuning* model IndoBERT untuk tugas klasifikasi emosi multi-kelas pada teks informal bahasa Indonesia yang berasal dari media sosial Twitter. Dengan membangun sebuah dataset

kustom yang seimbang melalui *scraping* bertarget dan menerapkan proses pelabelan yang tervalidasi secara statistik, model yang dikembangkan mampu mengatasi tantangan linguistik seperti penggunaan bahasa gaul dan singkatan. Hasil evaluasi pada data uji membuktikan bahwa model yang di-*fine-tuning* mencapai kinerja yang sangat tinggi dengan akurasi sebesar 94% dan *F1-Score* rata-rata tertimbang 0.94. Analisis kurva pembelajaran juga mengonfirmasi bahwa model yang dilatih sehat secara teknis dan tidak mengalami *overfitting*, sehingga memiliki kemampuan generalisasi yang andal pada data baru. Kinerja yang kuat dan seimbang di kelima kelas emosi (marah, takut, senang, cinta, dan sedih) menegaskan bahwa IndoBERT, ketika diadaptasi dengan benar, merupakan solusi yang sangat kuat untuk tugas-tugas *Natural Language Processing* pada domain teks informal bahasa Indonesia. Kinerja model yang tinggi ini menunjukkan potensinya sebagai komponen inti yang andal untuk berbagai aplikasi praktis yang bergantung pada pemahaman emosi pengguna. Selain itu, model ini dapat diaplikasikan dalam analisis sentimen publik, pemantauan citra merek, atau layanan pelanggan cerdas. Meskipun model IndoBERT menunjukkan kinerja yang sangat tinggi, penelitian ini memiliki beberapa keterbatasan yang dapat menjadi arahan untuk penelitian selanjutnya. Keterbatasan utama adalah fokus penelitian yang hanya mengevaluasi satu arsitektur model (IndoBERT) tanpa melakukan perbandingan formal dengan model *baseline* lain, seperti *machine learning* tradisional (misalnya, SVM atau Naïve Bayes) atau arsitektur *deep learning* alternatif (seperti LSTM/Bi-LSTM). Selain itu, seperti yang telah dibahas di metodologi, validasi model dapat diperkuat dengan menggunakan pembagian data 3-lapis (latih/validasi/uji) yang terpisah secara ketat atau menerapkan metode *k-fold cross-validation*. Penelitian selanjutnya juga dapat dilakukan dengan menambahkan kelas emosi yang lebih beragam (misalnya, terkejut atau jijik) atau mengembangkan teknik khusus untuk mendeteksi ekspresi emosi yang lebih kompleks seperti sarkasme dan ironi.

REFERENCES

[1] A. Hasnining and Y. Hazriani, "Text Mining Untuk Klasifikasi Emosi Pengguna Media Sosial Degan Algoritma Naïve Bayes," *ARTHA Technological Journal*, vol. 7, pp. 57–67, 2023, doi: 10.33857/patj.v7i1.671.

[2] M. Mustak, H. Hallikainen, T. Laukkanen, L. Plé, L. D. Hollebeek, and M. Aleem, "Using Machine Learning to Develop Customer Insights From User-Generated Content," *Journal of Retailing and Consumer Services*, vol. 81, p. 104034, Nov. 2024, doi: 10.1016/j.jretconser.2024.104034.

[3] M. I. Raif, N. N. Hidayati, and T. Matulatan, "Otomatisasi Pendeteksi Kata Baku dan Tidak Baku pada Data Twitter Berbasis KBBI," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 2, pp. 337–348, 2024, doi: 10.25126/jtiik.20241127404.

[4] M. I. Maulana, M. Fikry, S. Agustian, S. Ramadhani, and others, "Analisis Sentimen Ulasan Aplikasi Indodax Pada Google Play Store Dengan Algoritma Random Forest," *Bulletin of Computer Science Research*, vol. 5, no. 4, pp. 564–572, 2025, doi: 10.47065/bulletincsr.v5i4.626.

[5] D. E. Sondakh, R. C. Maringka, F. P. Ayorbaba, J. S. Mangi, and S. R. Pungus, "Emotion Mining User Review of the BRImo Mobile Banking Application Using the Decision Tree Algorithm," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 3, pp. 350–355, 2023, doi: 10.32736/sisfokom.v12i3.1721.

[6] J. Husna and M. D. A. Widirahayu, "Pemodelan Topik dan Analisis Sentimen pada 'Voices of History: 50 Iconic Speeches' Menggunakan Pendekatan Natural Language Processing," *Jurnal Ilmiah Manajemen Informasi dan Komunikasi*, vol. 8, no. 1, pp. 15–24, 2024, doi: 10.56873/jimik.v8i1.330.

[7] D. E. Putro, D. Juarsa, B. P. P. Hermana, B. Bagastian, and H. Sulistiani, "Analisis Sentimen Publik terhadap 'Save Raja Ampat'di Media Sosial Menggunakan Model IndoBERT," *Bulletin of Computer Science Research*, vol. 5, no. 5, pp. 1067–1075, 2025, doi: 10.47065/bulletincsr.v5i5.621.

[8] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The Hybrid of Bert and Deep Learning Models for Indonesian Sentiment Analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 591–602, 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.

[9] M. F. Kono, I. N. Fajri, and Y. Pristyanto, "Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2616–2628, 2025, doi: 10.30871/jaic.v9i5.10537.

[10] A. Safira and F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *ZONAsi: Jurnal Sistem Informasi*, vol. 5, no. 1, pp. 59–70, 2023, doi: 10.31849/zn.v5i1.12856.

[11] S. Mulyani, S. A. Thamrin, and S. Siswanto, "Analisis Sentimen Masyarakat pada Kebijakan Vaksinasi Covid-19 di Twitter Menggunakan Metode Mesin Vektor Pendukung dengan Kernel Radial Basis Function Berbasis Fitur Leksikon," *Jambura Journal of Probability and Statistics*, vol. 3, no. 2, pp. 110–119, 2022, doi: 10.34312/jjps.v3i2.16663.

[12] Y. Romadhoni and K. F. H. Holle, "Analisis Sentimen Terhadap PERMENDIKBUD No. 30 pada Media Sosial Twitter Menggunakan Metode Naive Bayes dan LSTM," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 118–124, 2022, doi: 10.30591/jpit.v7i2.3191.

[13] A. Hasiholan, I. Cholissodin, and N. Yudistira, "Analisis Sentimen Tweet Covid-19 Varian Omicron pada Platform Media Sosial Twitter menggunakan Metode LSTM berbasis Multi Fungsi Aktivasi dan GLOVE," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 6, no. 10, pp. 4653–4661, 2022.

[14] P. T. Astuti, "Sentiment Analysis Sudut Pandang Generasi Z terhadap Keterwakilan Kepemimpinan Muda Pilpres di Twitter Menggunakan ID Convolutional Neural Network," *Djtechno: Jurnal Teknologi Informasi*, vol. 5, no. 2, pp. 275–288, 2024, doi: 10.46576/djtechno.v5i2.4653.

[15] C. J. L. Tobing, I. G. N. L. Wijayakusuma, and L. P. I. Harini, "Perbandingan Kinerja IndoBERT dan MBERT untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia," *JST (Jurnal Sains dan Teknologi)*, vol. 14, no. 1, pp. 114–123, 2025, doi: 10.23887/jstundiksha.v14i1.92126.

[16] C. Ramadhan, V. Atina, and H. Permatasari, "Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2025, pp. 110–118. doi: 10.47701/v1r9ka69.

[17] P. Sayarizki, H. Nurrahmi, and others, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Indonesian Journal on Computing (Indo-JC)*, vol. 9, no. 2, pp. 61–72, 2024, doi: 10.34818/INDOJC.2024.9.2.934.
[18] M. R. Rabbani, H. M. Manik, and T. Hestirianoto, "Klasifikasi Gelembung Gas Menggunakan Multibeam Echosounder dan Machine Learning," *Jurnal Kelautan Tropis*, vol. 28, no. 2, pp. 247–254, 2025, doi: 10.14710/jkt.v28i2.26778.
[19] A. M. Andrés and M. Á. Hernández, "Estimators of Various Kappa Coefficients Based on the Unbiased Estimator of the Expected Index of Agreements," *Adv Data Anal Classif*, vol. 19, no. 1, pp. 177–207, 2025, doi: 10.1007/s11634-024-00581-x.
[20] L. Yang, X. Zhou, J. Fan, X. Xie, and S. Zhu, "Can Bidirectional Encoder Become the Ultimate Winner for Downstream Applications of Foundation Models?," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 2024, pp. 526–534. doi: 10.1109/FLLM63129.2024.10852511.
[21] J. Sun, Y. Liu, J. Cui, and H. He, "Deep Learning-based Methods for Natural Hazard Named Entity Recognition," *Sci Rep*, vol. 12, p. 4598, Jun. 2022, doi: 10.1038/s41598-022-08667-2.
[22] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," in *Proceedings of the Second Workshop in South East Asian Language Processing*, D. Wijaya, A. F. Aji, C. Vania, G. I. Winata, and A. Purwarianti, Eds., Association for Computational Linguistics, 2025, pp. 10–26.
[23] N. P. I. Maharani, A. Purwarianti, Y. Yustiawan, and F. C. Rochim, "Domain-Specific Language Model Post-Training for Indonesian Financial NLP," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346625.
[24] I. Alam, G. Nabiilah, E. S. Purwanto, and M. F. Hidayat, "Indonesian Multilabel Classification Using IndoBERT Embedding and MBERT Classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 1071–1078, Jun. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
[25] G. M. Foody, "Challenges in the Real World Use of Classification Accuracy Metrics: From Recall and Precision to the Matthews Correlation Coefficient," *PLoS One*, vol. 18, no. 10, p. e0291908, 2023, doi: 10.1371/journal.pone.0291908.