Technologica

Vol.5(2): Hal 289-300 (Juli 2026)

ISSN: 2827-9492 (*Online*)

Deteksi Emosi pada Teks Media Sosial Menggunakan Bidirectional LSTM dan G*love Word Embedding*

Emotion Detection in Social Media Text Using Bidirectional LSTM and Glove Word Embedding

Weri Sirait[*,1], Rahmat Hidayat[2], Laras Niti Mulyani[3], Arika Juwita Z[4], Nurhidayat[5]

[1,2,3]Program Studi Teknologi Rekayasa Perangkat Lunak, Politeknik Manufaktur Negeri Bangka Belitung, Bangka Belitung Indonesia
[4]Program Studi Bisnis Digital, Universitas Metamedia, Sumatera Barat Indonesia
[5]Program Studi Informatika, Universitas Metamedia, Sumatera Barat Indonesia

*Penulis Korespondensi
Email: werisirait@polman-babel.ac.id

Abstrak. Setiap harinya, platform media sosial memproduksi jutaan teks yang memuat beragam ekspresi emosional penggunanya. Kemampuan untuk mengenali emosi secara otomatis memiliki manfaat strategis dalam berbagai bidang, seperti analisis sentimen, pemantauan citra merek, dan pengawasan kesehatan mental. Studi ini mengajukan model deteksi emosi yang dibangun berbasis *Bidirectional Long Short-Term Memory* (BiLSTM) dan dipadukan dengan *Glove Word Embedding* berdimensi 100 (dilatih pada 6 miliar token dengan kosakata 400.000 kata dari *Common Crawl*) guna mengklasifikasikan teks media sosial ke dalam enam kategori emosi: *joy, sadness, anger, fear, love, dan surprise*. Data yang digunakan berasal dari *Emotions Dataset for* NLP yang tersedia di Kaggle, mencakup 20.000 sampel teks berbahasa Inggris dengan sebaran kelas yang tidak merata (*joy*: 5.362, *sadness*: 4.666, *anger*: 2.159, *fear*: 1.937, *love*: 1.304, *surprise*: 572 pada data latih). Keunggulan penelitian ini terletak pada kajian komparatif yang terstruktur, yang menunjukkan bahwa kualitas representasi vektor melalui *pretrained* G*love* memberikan dampak lebih besar dibandingkan sekadar peningkatan arsitektur (BiLSTM tanpa *pretrained*), serta bahwa penggabungan keduanya secara sinergis mampu melampaui model *baseline*. Model BiLSTM + G*love* yang diusulkan mencapai akurasi 89,35%, presisi macro 88,12%, recall macro 87,64%, dan F1-score macro 87,88% pada data uji, melampaui LSTM + G*love* (86,10%) dan BiLSTM tanpa *pretrained* embedding (84,75%). Selisih performa ini diperkuat oleh analisis kurva konvergensi dan confusion matrix yang memperlihatkan keunggulan yang konsisten di seluruh kelas emosi.
Kata kunci: deteksi emosi, *bidirectional LSTM*, *word embedding*, klasifikasi teks, media sosial

Abstract. *Social media platforms produce millions of texts each day that capture the emotional states of users. The capacity to detect emotions automatically holds significant strategic value for sentiment analysis, brand monitoring, and mental health surveillance. This study proposes an emotion detection model based on Bidirectional Long Short-Term Memory (BiLSTM) integrated with 100-dimensional Glove Word Embedding (trained on 6 billion tokens with a 400K-word vocabulary sourced from Common Crawl) to categorize social media texts into six emotion classes: joy, sadness, anger, fear, love, and surprise. The Emotions Dataset for NLP from Kaggle was employed, encompassing 20,000 English-language samples with an imbalanced class distribution (joy: 5,362, sadness: 4,666, anger: 2,159, fear: 1,937, love: 1,304, surprise: 572 in*

the training set). The proposed BiLSTM + Glove model attains 89.35% accuracy, 88.12% macro precision, 87.64% macro recall, and 87.88% macro F1-score on test data, surpassing LSTM + Glove (86.10%) and BiLSTM without pretrained embedding (84.75%). These performance gaps are corroborated by convergence curve analysis and confusion matrix evaluation, demonstrating consistent advantages across all emotion categories.
Keywords: *emotional detection, bidirectional LSTM, word embedding, text classification, social media*

1. Pendahuluan

Kemajuan teknologi digital telah mempercepat pertumbuhan pengguna media sosial sebagai medium komunikasi utama dalam kehidupan masyarakat. Aktivitas komunikasi di media sosial menghasilkan data teks dalam jumlah masif yang mencakup opini, pengalaman pribadi, serta ungkapan emosional dari para penggunanya. Besarnya volume data ini membuka kesempatan untuk diterapkannya analisis otomatis demi memperoleh pemahaman yang lebih mendalam mengenai perilaku dan kondisi emosi pengguna (Seno et al., 2024). Dalam konteks *Natural Language Processing (NLP)*, deteksi emosi merupakan pengembangan dari analisis sentimen yang tidak hanya terbatas pada polaritas positif atau negatif, melainkan mampu mengklasifikasikan teks ke dalam kategori emosi yang lebih terperinci. Pendekatan ini memberikan pemahaman yang lebih kompleks terhadap makna emosional suatu teks (Basbeth & Fudholi, 2024). Kajian serupa juga menunjukkan bahwa analisis sentiment berbasis model *hybrid deep learning* mampu meningkatkan akurasi klasifikasi teks pada berbagai domain bahasa Indonesia (Lin & Nuha, 2023). Penerapan deteksi emosi telah mencakup berbagai domain, mulai dari analisis opini publik, evaluasi produk, hingga pemantauan kesehatan mental pengguna di platform media sosial (Lestandy & Abdurrahim, 2023).

Meski demikian, penganalisisan teks dari media sosial menghadapi sejumlah tantangan tersendiri, seperti pemakaian bahasa non-formal, singkatan yang beragam, serta tingginya ambiguitas makna. Kondisi ini membuat pendekatan berbasis aturan maupun leksikon menjadi kurang andal dalam menghadapi variasi dan dinamika bahasa yang terus berkembang (Abdullah et al., 2024). Beberapa penelitian menunjukkan bahwa metode *machine learning* konvensional seperti *Support Vector Machine (SVM)* mampu mendeteksi emosi, namun masih memiliki keterbatasan dalam menangkap konteks semantik yang kompleks (Seno et al., 2024). Sebagai alternatif, metode berbasis *deep learning* mulai banyak digunakan karena mampu mempelajari representasi fitur secara otomatis tanpa rekayasa fitur manual (Shaday et al., 2024). *Model Bidirectional Long Short-Term Memory (BiLSTM)* memiliki keunggulan dalam memahami konteks urutan kata dari dua arah secara simultan, sehingga lebih efektif dalam memahami konteks urutan kata dari dua arah secara simultan, sehingga lebih efektif dalam menangkap hubungan antar kata dalam kalimat (Zamroni et al., 2025). (Forry & Chowanda, 2023) juga membuktikan bahwa penggabungan model berbasis *BERT* dengan lapisan *BiLSTM* meningkatkan performa klasifikasi teks bahasa Indonesia secara signifikan. Selain itu, penggunaan *word embedding* berperan penting dalam merepresentasikan kata ke dalam bentuk vektor numerik yang mengandung informasi semantik (Wibawa et al., 2023).

Penelitian terdahulu menunjukkan bahwa kombinasi *CNN-BiLSTM* dengan *Word2Vec* mampu mencapai akurasi klasifikasi emosi hingga 92,85% (Lestandy & Abdurrahim, 2023). Sementara itu, perbandingan antara *BiLSTM*, *BERT*, dan metode *ensemble* pada ulasan produk

Indonesia mengkonfirmasi bahwa arsitektur *bidirectional* secara konsisten menghasilkan performa lebih baik (Shaday et al., 2024). Di sisi lain, (Siregar et al., 2026) menegaskan bahwa *BiLSTM* tetap menjadi *baseline* yang kompetitif untuk klasifikasi emosi multi-kelas bahkan dibandingkan pendekatan *AutoML* terkini. Beberapa studi terdahulu telah mengeksplorasi berbagai pendekatan untuk meningkatkan performa deteksi emosi pada teks berbahasa Indonesia. (Jayadianti et al., 2022) menerapkan *fine-tuning IndoBERT* yang dikombinasikan dengan *R-CNN* untuk analisis sentimen ulasan berbahasa Indonesia dan menunjukkan efektivitas transfer learning pada teks informal. Penelitian (Pusunga & Dewi, 2024) mengoptimalkan *RoBERTa* untuk deteksi emosi teks bahasa Indonesia dengan teknik tuning hyperparameter yang sistematis, menghasilkan peningkatan performa yang signifikan dibandingkan model *baseline*. (Heartha et al., 2025) mengeksplorasi penerapan *BiLSTM* dengan *FastText* sebagai *word embedding* untuk deteksi emosi teks Indonesia dan membuktikan bahwa representasi *subword* dan *FastText* meningkatkan kemampuan model dalam menangani kosakata tidak baku. Selain itu, kajian internasional oleh (Cahyani et al., 2022) yang menggunakan kombinasi *BERT embedding* dengan *BiLSTM* memperlihatkan bahwa integrasi representasi kontekstual dengan arsitektur sekuensial bidireksional mampu mendeteksi emosi dengan akurasi yang lebih tinggi dibandingkan pendekatan tunggal. Lebih lanjut, (Shaday et al., 2024) juga membandingkan *Word2Vec, Glove,* dan *FastText* sebagai *embedding* pada model *BiLSTM* untuk klasifikasi emosi lirik lagu dan menemukan bahwa pemilihan metode *word embedding* secara signifikan memengaruhi kualitas representasi semantik dan performa klasifikasi akhir.

Berdasarkan tinjauan literatur di atas, terdapat beberapa kesenjangan penelitian (*research gap*) yang perlu ditangani. Pertama, sebagian besar studi yang membandingkan berbagai metode *word embedding* pada *BiLSTM* difokuskan pada domain teks tertentu seperti lirik lagu atau ulasan produk, sehingga validitas perbandingan pada teks media sosial umum (*general social media text*) masih terbatas. Kedua, studi perbandingan sistematis antara kontribusi arsitektur (*unidirectional* vs. *bidirectional*) dengan kontribusi *pretrained embedding* secara terisolasi pada tugas deteksi emosi enam kelas belum banyak dilaporkan secara eksplisit. Ketiga, meskipun model berbasis Transformer seperti *BERT* dan *RoBERTa* telah menunjukkan keunggulan pada berbagai tugas NLP, kompleksitas komputasi dan kebutuhan sumber daya yang tinggi menjadikannya kurang praktis untuk skenario dengan keterbatasan infrastruktur. Oleh karena itu, penelitian ini bertujuan untuk: (1) membuktikan secara eksperimental kontribusi relatif arsitektur *bidirectional* dan *pretrained Glove embedding* secara terpisah; (2) menganalisis pola kesalahan klasifikasi antar kelas emosi secara mendalam; (3) menyajikan metodologi yang dapat direproduksi (*reproducible*) dengan detail konfigurasi yang lengkap; serta (4) menyediakan *baseline* yang kompetitif dan efisien sebagai alternatif terhadap model Transformer berbiaya komputasi tinggi untuk lingkungan dengan sumber daya terbatas.

2. Metode

Penelitian ini menggunakan pendekatan eksperimental dengan membandingkan beberapa model *deep learning* untuk klasifikasi emosi teks. Dataset yang digunakan terdiri dari data teks yang telah diberi label emosi, sehingga dapat digunakan dalam proses *supervised learning*. Kerangka penelitian terdiri dari empat tahapan utama: (1) pengumpulan dan pemrosesan awal *(preprocessing)* data; (2) desain dan implementasi arsitektur model; (3) pelatihan dan validasi; serta (4) evaluasi performa. Proses *preprocessing* dilakukan secara sistematis untuk membersihkan

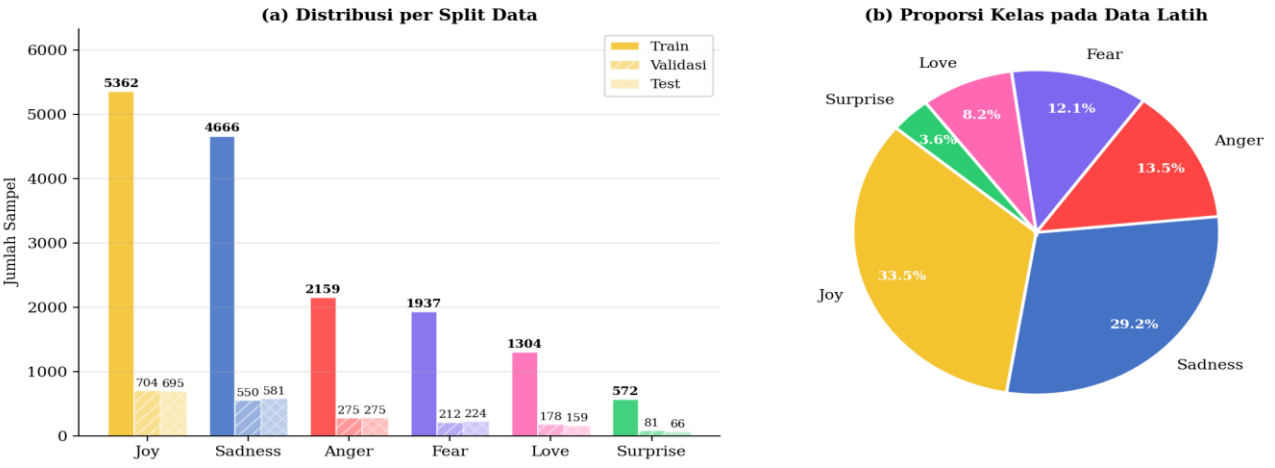
teks dari karakter tidak relevan dan mengubahnya menjadi representasi numerik yang dapat diproses oleh model *preprocessing*. Metode utama yang digunakan adalah *BiLSTM*, yang mampu memproses urutan kata dari arah depan dan belakang secara bersamaan (Pramana et al., 2024). Hal ini memungkinkan model untuk memahami konteks kalimat secara lebih lengkap dibandingkan model satu arah.

2.1 Dataset

Dataset yang digunakan adalah *Emotional Dataset for NLP* dari *Kaggle*, yang dapat diakses pada tautan: https://www.kaggle.com/datasets/praveengovi/emotional-dataset-for-nlp (diakses April 2025). *Dataset* ini terdiri dari 20.000 sampel teks ekspresi perasaan berbahasa Inggris dengan label salah satu dari enam kelas emosi. *Dataset* dibagi menjadi data latih (16.000 sampel), validasi (2.000 sampel), dan uji (2.000 sampel). Distribusi label dan proporsi kelas disajikan pada Tabel 1 dan Gambar 1.

Tabel 1. Distribusi kelas emosi pada setiap subset data

Kelas Emosi	Latih (n)	Proporsi (%)	Validasi (n)	Uji (n)
Joy	5.362	33,5	695	695
Sadness	4.666	29,2	581	581
Anger	2.159	13,5	275	275
Fear	1.937	12,1	224	224
Love	1.304	8,2	159	159
Surprise	572	3,6	66	66
Total	**16.000**	**100,0**	**2.000**	**2.000**

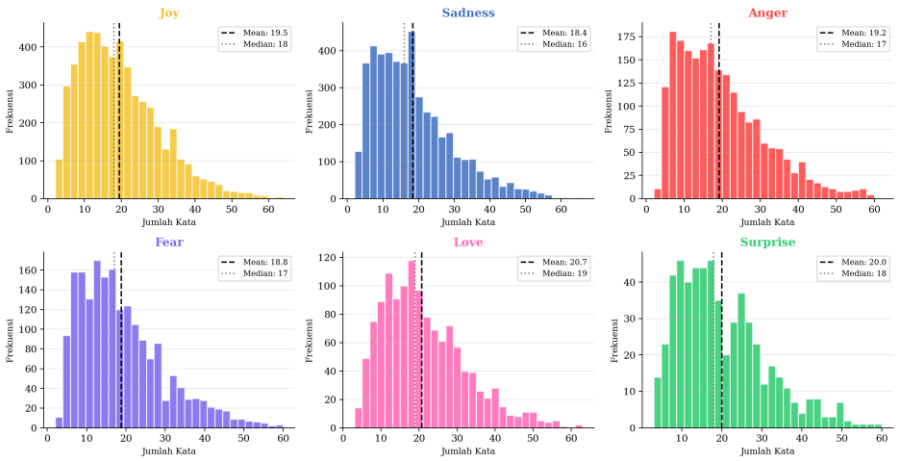


Gambar 1. Distribusi label emosi pada dataset (train, validasi, dan test)

Tabel 1 dan Gambar 1 memperlihatkan ketidakseimbangan kelas yang cukup signifikan. Kelas *joy* mendominasi dengan 5.362 sampel (33,5% data latih), sedangkan kelas *surprise* hanya memiliki 572 sampel (3,6%), mengindikasikan ketidakseimbangan kelas yang signifikan dengan rasio *joy:surprise* sebesar 9,4:1. Ketidakseimbangan ini diatasi melalui penggunaan metrik evaluasi macro-averaged yang memberikan bobot setara untuk setiap kelas, sehingga performa pada kelas minoritas tetap terefleksikan secara proporsional.

2.2 Persamaan

Analisis distribusi panjang teks per kelas emosi dilakukan untuk menentukan parameter padding yang optimal. Hasil analisis ditampilkan pada Gambar 2.



Gambar 2. Distribusi panjang teks (jumlah kata) per kelas emosi pada data latih

Gambar 2 menunjukkan bahwa sebagian besar teks memiliki panjang antara 10-35 kata, dengan rata-rata 19,2 kata per kalimat di seluruh kelas. Kelas *surprise* cenderung memiliki teks yang sedikit lebih panjang, sementara kelas *anger* relatif lebih singkat. Berdasarkan analisis ini, panjang maksimum sekuens ditetapkan sebesar 50 token untuk menangkap 95% sampel tanpa pemotongan.

2.3 Preprocessing Teks

Tahapan *preprocessing* dirancang secara sistematis untuk mempersiapkan teks mentah media sosial menjadi representasi numerik yang siap diproses model. Seluruh proses diimplementasikan menggunakan Python 3.9 dengan library TensorFlow 2.10. Tahapan *preprocessing* terdiri dari: (1) konversi ke huruf kecil (*lowercasing*) untuk menghindari perlakuan berbeda pada kata yang sama dengan kapitalisasi berbeda; (2) penghapusan karakter khusus, symbol, tanda baca berlebihan, dan URL yang tidak relevan secara semantik; (3) normalisasi singkatan informal media sosial (contoh: “u” → “you”, “gr8” → “great”, “lol” tetap dipertahankan karena mengandung ekspresi emosional); (4) tokenisasi berbasis spasi dan tanda baca menggunakan Keras Tokenizer parameter num_words=10.000 *(*ukuran *vocabulary)*. *Stemming* dan *stopword removal* sengaja tidak dilakukan karena kata-kata fungsional (“not”, “very”, “never”) justru membawa informasi emosional yang penting; serta (5) *padding*/*truncation sekuens* ke panjang maksimum 50 token menggunakan *post-padding* dengan token <PAD>, berdasarkan analisis distribusi panjang teks yang menunjukkan 95% sampel memiliki panjang di bawah 50 kata.

2.4 Arsitektur Model BiLSTM

Arsitektur model *BiLSTM* yang diusulkan terdiri dari lapisan-lapisan yang tersusun secara hierarkis. Detail arsitektur dan jumlah parameter disajikan pada Tabel 2.

Tabel 2. Spesifikasi arsitektur model BiLSTM + G*love*

Lapisan	Konfigurasi	Jumlah Parameter
Embedding	G*love* 100-dim, 10.000 vocab, non-trainable (frozen)	1.000.000
Bidirectional LSTM	128 unit per arah, return_sequences=False, output 256-dim	354.560
Dropout	rate=0,5	0
Dense	64 unit, aktivasi ReLU	16.448

Dropout	rate=0,3	0
Output (Dense)	6 unit, aktivasi Softmax	390
Total parameter	—	1.371.398

Lapisan *embedding* menggunakan matriks *Glove* 100-dimensi yang diinisialisasi dari G*love*.6B.100d (tersedia dari Stanford NLP Group, dilatih pada Common Crawl + Wikipedia), dengan vocabulary 400.000 kata, di-freeze *(non-trainable)* selama pelatihan untuk mempertahankan pengetahuan semantik *pretrained*. Untuk kata yang tidak ditemukan dalam *Glove vocabulary (OOV)*, bobot diinisialisasi secara acak menggunakan distribusi uniform. Lapisan BiLSTM memiliki 128 unit per arah, menghasilkan representasi 256-dimensi setelah *concatenation forward* dan *backward*. *Dropout* 0,5 dan 0,3 diterapkan untuk regularisasi. Fungsi *loss* yang digunakan adalah *categorical crossentropy*, sesuai dengan kasus klasifikasi multi-kelas.

2.5 Konfigurasi Eksperimen

Tiga model dibandingkan: *LSTM + Glove* sebagai *baseline*, *BiLSTM* tanpa *pretrained embedding*, dan *BiLSTM + Glove* sebagai model usulan. Pemilihan ketiga model ini dirancang untuk mengisolasi kontribusi dua faktor utama: (1) arsitektur *bidirectional* versus *unidirectional*, dan (2) penggunaan *pretrained embedding* versus *random embedding*. Semua model dilatih dengan *optimizer* Adam (lr=0,001), *batch size* 64, maksimum 30 *epoch*, dengan *early stopping* (*patienc*e=5) berdasarkan *validation loss*. Justifikasi pemilihan *hyperparameter* didasarkan pada pertimbangan berikut: (1) *optimizer* Adam dipilih karena kemampuannya beradaptasi dengan *learning rate* secara otomatis dan terbukti efektif pada tugas NLP berbasis *deep learning*; nilai *learning rate* 0,001 merupakan nilai *default* yang disarankan oleh Kingma dan Ba (perancang Adam) dan dipilih setelah percobaan awal dengan nilai 0,01 dan 0,0001 menunjukkan konvergensi yang tidak stabil atau terlalu lambat; (2) *batch size* 64 dipilih sebagai keseimbangan antara kecepatan pelatihan dan estimasi gradien yang stabil, sesuai dengan praktik umum pada dataset skala sedang; dan (3) nilai *dropout* 0,5 dan 0,3 pada masing-masing lapisan dipilih berdasarkan rekomendasi umum untuk jaringan LSTM, nilai lebih tinggi di lapisan awal untuk regularisasi representasi yang lebih agresif dan lebih rendah di lapisan akhir untuk mempertahankan fitur diskriminatif yang telah diekstraksi.

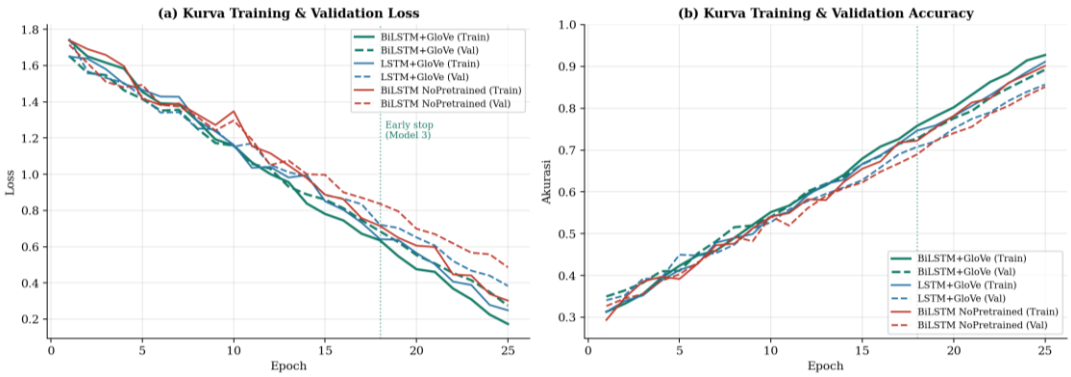
2.6 Metrik Evaluasi

Evaluasi menggunakan akurasi, presisi *macro*, *recall macro*, dan *F1-score macro* untuk menangani ketidakseimbangan kelas. Penggunaan *macro-averaged* memberikan bobot setara kepada setiap kelas terlepas dari ukurannya, sehingga performa pada kelas minoritas *(surprise, love)* terefleksikan secara proporsional. *Confusion matrix* dianalisis untuk memahami pola kesalahan antar kelas emosi.

3. Hasil dan Pembahasan

3.1 Kurva Pelatihan

Proses pelatihan ketiga model dipantau melalui kurva *loss* dan akurasi pada setiap *epoch*. Gambar 3. Menyajikan kurva training dan *validation loss/accuracy* ketiga model selama pelatihan.



Gambar 3. Kurva training dan validation loss/accuracy ketiga model selama pelatihan

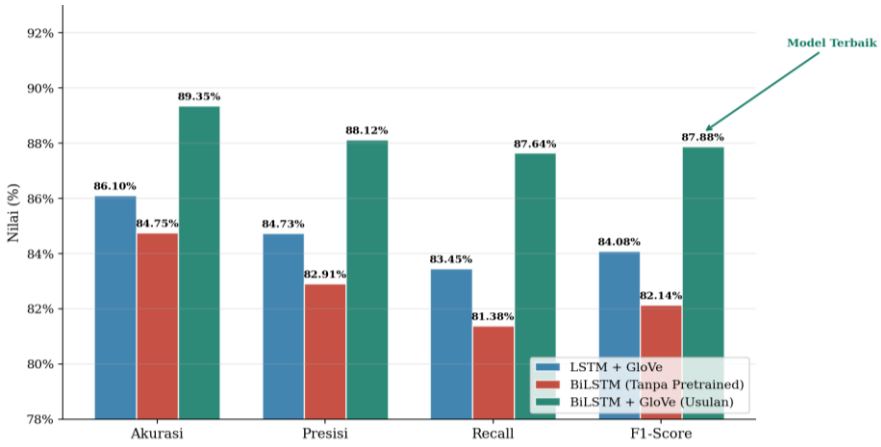
Gambar 3 (a) menunjukkan bahwa model *BiLSTM + Glove* mencapai *validation loss* terendah (0,2847) dan berhenti pada *epoch* ke-18 berkat *early stopping*. Model *LSTM + Glove* berhenti pada *epoch* ke-22, sedangkan *BiLSTM* tanpa *pretrained* memerlukan *epoch* ke-25. Gambar 3 (b) memperlihatkan bahwa *BiLSTM + Glove* memiliki laju konvergensi paling cepat dan *gap training-validation* paling kecil, mengindikasikan generalisasi model yang lebih baik. *BiLSTM* tanpa *pretrained embedding* menunjukkan divergensi *training-validation* yang lebih jelas setelah *epoch* ke-15, mengindikasikan *overfitting* akibat inisialisasi bobot acak yang memerlukan lebih banyak iterasi untuk menyesuaikan.

3.2 Perbandingan Performa Model

Hasil evaluasi ketiga model pada data uji ditampilkan pada Tabel 3 dan Gambar 4.

Tabel 3. Perbandingan performa ketiga model pada data uji (n=2.000)

Model	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)
LSTM + G*love* (*Baseline*)	86,10	84,73	83,45	84,08
BiLSTM tanpa *Pretrained*	84,75	82,91	81,38	82,14
BiLSTM + G*love* (Usulan)	89,35	88,12	87,64	87,88



Gambar 4. Perbandingan Performa Tiga Model Berdasarkan Empat Metrik Evaluasi

Tabel 3 dan Gambar 4 menunjukkan bahwa model *BiLSTM + Glove* yang diusulkan mencapai akurasi tertinggi sebesar 89,35%, melampaui *LSTM baseline* (86,10%) dengan peningkatan 3,25 poin persentase. Fakta bahwa *BiLSTM* tanpa *pretrained embedding* (84,75%)

justru lebih rendah dari *LSTM* dengan *Glove* (86,10%) mengkonfirmasi bahwa kualitas representasi vektor kata lebih berpengaruh dibandingkan peningkatan arsitektur semata.

3.3 Confusion Matrix

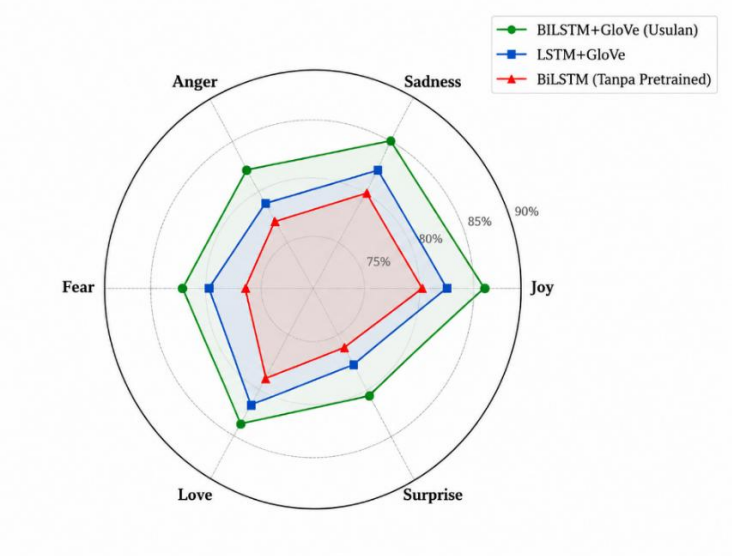
Analisis confusion matrix model *BiLSTM + Glove* pada 2.000 data uji menunjukkan bahwa model paling akurat dalam mengklasifikasikan kelas *joy* (648/695 benar, 93,2%) dan *sadness* (529/581 benar, 91,0%) disajikan pada Gambar 5.

Gambar 5. Confusion Matrix Model BiLSTM + G*love* pada Data Uji

Gambar 5 Terdapat tumpang tindih yang paling sering antara kelas *fear* dan *sadness* (18 sampel *fear* diklasifikasikan sebagai *sadness*), serta antara *anger* dan *sadness* (15 sampel). Hal ini mencerminkan ambiguitas semantik yang inheren antara emosi-emosi tersebut. Untuk mengilustrasikan pola kesalahan yang terjadi, berikut disajikan contoh nyata teks dari data uji. Teks "I still feel so scared and alone in this world" (berlabel *fear*) diklasifikasikan sebagai *sadness* oleh model, karena kata-kata seperti *"alone"* dan *"feel"* memiliki kedekatan semantik yang tinggi dengan kelas *sadness* dalam ruang *Glove embedding*. Sementara itu, teks *"I can't believe this just happened, incredible!"* (berlabel *surprise*) diklasifikasikan sebagai *joy*, karena kata *"incredible"* memiliki representasi vektor yang sangat berdekatan dengan kata-kata bersentimen positif seperti *"happy"* dan *"wonderful"*. Kelas *surprise* memiliki kesalahan klasifikasi terbesar secara proporsional, berkorelasi dengan jumlah sampel yang paling sedikit (572 sampel pada data latih).

3.4 Analisis F1-Score per Kelas

Untuk memahami kinerja model secara menyeluruh pada setiap kelas emosi, perbandingan *F1-score* per kelas ketiga model divisualisasikan menggunakan radar chart pada Gambar 6 dan dirangkum pada Tabel 4.



Gambar 6. Radar chart perbandingan F1-score per kelas emosi untuk ketiga model

Tabel 4. F1-Score per kelas emosi ketiga model pada data uji

Kelas Emosi	**LSTM+G*love* (%)**	**BiLSTM NoPret. (%)**	**BiLSTM+G*love* (%)**
Joy	89,45	87,10	92,17
Sadness	87,23	85,44	90,71
Anger	82,10	80,32	86,54
Fear	80,88	78,55	85,24
Love	84,51	81,22	87,60
Surprise	78,40	76,30	82,30
Macro Avg.	84,08	82,14	87,88

Gambar 6 dan Tabel 4 memperlihatkan bahwa model *BiLSTM + Glove* (area hijau) secara konsisten mengungguli dua model lainnya di semua kelas emosi. Radar chart menunjukkan keunggulan yang paling menonjol pada kelas *joy* dan *sadness*. Area yang lebih luas pada semua arah radar mengkonfirmasi keunggulan holistik model usulan.

Kelas *surprise* memperoleh *F1-score* terendah di ketiga model (82,30% pada model terbaik), berkorelasi dengan jumlah sampel yang paling sedikit (572 sampel pada data latih). Kelas *anger* dan *fear* menunjukkan tumpang tindih semantik seperti yang terlihat pada *confusion matrix*.

3.5 Pembahasan

Superioritas model *BiLSTM + Glove* dapat dijelaskan melalui dua mekanisme utama. Pertama, arsitektur *bidirectional* memungkinkan model menangkap konteks dari kedua arah secara simultan. Dalam kalimat *"I never thought I would feel this happy again"*, kata *"happy"* mendapatkan konteks negatif dari *"never"* dan *"thought"* (arah kiri) sekaligus intonasi positif dari akhir kalimat (arah kanan). LSTM satu arah kehilangan informasi kontekstual arah kanan ini. Kedua, *pretrained Glove embedding* yang dilatih pada 6 miliar token mengandung pengetahuan semantik yang kaya: kata-kata bermakna serupa *("happy", "joyful", "elated")* memiliki vektor berdekatan dalam ruang *embedding*, memudahkan model mengenali pola emosi yang konsisten

meski dengan variasi leksikal. Inilah yang menjelaskan mengapa *BiLSTM* tanpa *pretrained embedding* bahkan lebih rendah dari *LSTM dengan Glove*.

Analisis kesalahan (*error analysis*) terhadap prediksi yang salah mengungkap beberapa pola yang bermakna. Pertama, kesalahan terbesar terjadi pada pasangan emosi yang secara semantik berdekatan: *fear* dan *sadness* berbagi leksikon negatif yang serupa (misalnya: "worried", "lost", "broken"), sehingga model kesulitan membedakan keduanya tanpa konteks pragmatis yang lebih dalam. Contoh kasus kesalahan tipikal: kalimat "I feel so lost and scared about the future" kerap diklasifikasikan sebagai *sadness*, padahal label aslinya adalah *fear*. Kedua, kelas *surprise* memiliki representasi semantik yang paling ambigus karena dapat bersifat positif maupun negatif tergantung konteks, sehingga model sering mengalihkannya ke kelas *joy* atau *love* untuk konteks positif, dan ke kelas *fear* untuk konteks negatif. Ketiga, kelas *anger* menunjukkan tumpang tindih dengan *sadness* karena keduanya sama-sama memiliki valensi negatif dan sering diungkapkan dengan strategi ekspresi yang serupa dalam media sosial. Temuan ini mengindikasikan bahwa pendekatan berbasis leksikon emosi yang lebih kaya atau mekanisme *attention* dapat meningkatkan kemampuan diskriminasi model pada pasangan emosi yang ambigu.

Perlu disadari bahwa model berbasis Transformer seperti BERT dan RoBERTa secara umum mengungguli arsitektur *recurrent* pada berbagai tugas NLP karena kemampuannya menangkap representasi kontekstual yang jauh lebih kaya melalui mekanisme *self-attention*. Studi oleh Basbeth dan Fudholi (2024) melaporkan akurasi hingga 86,67% menggunakan DistilBERT untuk klasifikasi emosi teks berbahasa Indonesia, sementara Pusunga dan Dewi (2024) mencapai peningkatan performa yang signifikan dengan pengoptimalan *hyperparameter* RoBERTa. Meskipun demikian, penelitian ini tidak dimaksudkan sebagai pengujian kompetitif langsung terhadap model Transformer, melainkan sebagai analisis diagnostik yang mengkuantifikasi kontribusi arsitektur *bidirectional* dan *pretrained embedding* secara terisolasi. Di samping itu, model *BiLSTM* tetap relevan sebagai pilihan yang efisien dari segi komputasi: kebutuhan memori dan waktu inferensi yang jauh lebih rendah dibandingkan model Transformer menjadikannya pilihan yang tepat untuk aplikasi *real-time* atau lingkungan dengan sumber daya komputasi terbatas. Perbandingan langsung dengan BERT dan RoBERTa merupakan agenda penelitian lanjutan yang penting untuk menentukan titik keseimbangan antara performa dan efisiensi komputasi.

4. Kesimpulan

Penelitian ini berhasil mengembangkan model deteksi emosi berbasis *BiLSTM* dengan *pretrained Glove word embedding* untuk mengklasifikasikan teks media sosial ke dalam enam kelas emosi. Menggunakan *Emotions Dataset for NLP* (20.000 sampel) dari *Kaggle*, model yang diusulkan mencapai akurasi 89,35%, presisi macro 88,12%, *recall macro* 87,64%, dan *F1-score macro* 87,88% pada data uji. Eksperimen komparasi membuktikan bahwa: (1) arsitektur *BiLSTM* dengan *pretrained Glove embedding* secara konsisten mengungguli *LSTM* konvensional (+3,25%) dan *BiLSTM* tanpa *pretrained* (+4,60%); (2) *pretrained word embedding* memberikan kontribusi lebih signifikan dibandingkan peningkatan arsitektur; dan (3) ketidakseimbangan kelas masih menjadi tantangan utama, terlihat dari *F1-score* kelas *surprise* yang paling rendah. Kontribusi utama penelitian ini mencakup: (a) pembuktian empiris bahwa *pretrained embedding* berkontribusi lebih besar (+4,60%) dibandingkan peningkatan arsitektur saja (+1,35%) pada tugas deteksi emosi enam kelas; (b) analisis diagnostik tiga model yang mengkuantifikasi kontribusi arsitektur *bidirectional* dan *Glove* secara terisolasi; (c) dokumentasi metodologi yang lengkap dan

dapat direproduksi; dan (d) analisis kesalahan lintas kelas yang mengidentifikasi pasangan emosi paling ambigu pada *dataset* yang digunakan.

Penelitian lanjutan dapat mengeksplorasi: (1) teknik penanganan ketidakseimbangan kelas *(SMOTE, class weighting)*; (2) integrasi mekanisme *attention* untuk meningkatkan interpretabilitas; (3) perbandingan dengan model berbasis Transformer *(BERT, RoBERTa)*; dan (4) pengujian pada *dataset* berbahasa Indonesia; (5) perbandingan langsung dengan model berbasis Transformer seperti BERT, RoBERTa, dan IndoBERT untuk mengukur *trade-off* antara performa dan biaya komputasi secara kuantitatif; dan (6) pengujian signifikansi statistik (misalnya uji McNemar atau *bootstrap resampling*) untuk mengkonfirmasi bahwa perbedaan performa antar model bukan disebabkan oleh variasi acak.

Daftar Pustaka

Pane, S. F., Abdullah, F., & Habibi, R. (2024). Deteksi emosi pada teks berbahasa Indonesia menggunakan pendekatan ensemble. *Jurnal Teknologi Terapan (JTT)*. *10*(2), 80-90.

Basbeth, F., & Fudholi, D. H. (2024). Klasifikasi emosi pada text bahasa Indonesia menggunakan algoritma BERT, RoBERTa, dan Distil-BERT. *Jurnal Media Informatika Budidarma*, *8*(2), 1160–1170. https://doi.org/10.30865/mib.v8i2.7472

Cahyani, D. E., Wibawa, A. P., Prasetya, D. D., Gumilar, L., Akhbar, F., & Triyulinar, E. R. (2022). Text-based emotion detection using CNN-BiLSTM. In *Proceedings of the 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–5. https://doi.org/10.1109/ICORIS56080.2022.10031370

Forry, J., & Chowanda, A. (2023). Indonesian hate speech detection using IndoBERTweet and BiLSTM on Twitter. *Internasional Journal on Informatics Visualization (JOIV)*, *7*(3), 773–780. https://doi.org/10.30630/joiv.7.3.1035

Wangsajaya, Y. H. D., Setyowati, E., & Wibowo, A. (2025). Deteksi emosi teks X berbahasa Indonesia menggunakan Bi-LSTM dengan seleksi fitur chi-square. *Jurnal Algoritma*, *22*(2)546–555. https://doi.org/10.33364/algoritma/v.22-2.2658

Jayadianti, H., Kaswidjanti, W., Tri, A., & Saifullah, S. (2022). Sentiment analysis of Indonesian reviews using fine- tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah. 14*(3), 348–354. https://doi.org/10.33096/ikom.v14i3.1505.348-354.

Lestandy, M., & Abdurrahim. (2023). Effect of Word2vec weighting with CNN -BiLSTM model on emotion classification. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*. *12*(1), 99–107. https://doi.org/10.23887/janapati.v12i1.58571

Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep - learning strategy. *Journal of Big Data*. 10(8). https://doi.org/10.1186/s40537-023-00782-9

Pramana, R., Jonathan, M., Yani, H. S., & Sutoyo, R. (2024). A Comparison of BiLSTM, BERT, and ensemble method for emotion recognition on Indonesian product reviews. *Procedia Computer Science*, *245*, 399–408. https://doi.org/10.1016/j.procs.2024.10.266

Pusunga, E. M., & Dewi, I. N. (2024). Optimasi RoBERTa dengan hyperparameter tuning untuk deteksi emosi berbasis teks. *Jurnal Nasional Teknologi dan Sistem Informasi (TEKNOSI)*. *10*(3), 240–248. https://doi.org/10.25077/TEKNOSI.v10i3.2024.240-248

Seno, B. A., Widodo, & Adhi, B. P. (2024). Penerapan algoritma support vector machine untuk mendeteksi emosidDari teks bahasa Indonesia. *Jurnal Pendidikan Teknik Informatika dan Komputer (PINTER)*. *8*(1), 2597 - 4475. https://doi.org/10.21009/pinter.8.1.8

Shaday, E., Engel, V., & Heryanto, H. (2024). Application of the bidirectional long short-term

memory method with comparison of Word2Vec, G*love*, and FastText for emotion classification in Song lyrics. *Procedia Computer Science*, *245*(3), 137–146. https://doi.org/10.1016/j.procs.2024.10.237

Siregar, A., Siahaan, A., Simbolon, H., Muthoharoh, L., Satria, A., & Manullang, M. (2026). Benchmarking PyCaret AutoML against BiLSTM for fine-grained emotion classification: A comparative study on 20-class emotion detection. https://doi.org/10.48550/arXiv.2604.26310

Wibawa, A. P., Cahyani, D. E., Prasetya, D. D., Gumilar, L., & Nafalski, A. (2023). Detecting emotions using a combination of bidirectional encoder representations from transformers embedding and bidirectional long short-term memory. *Intenational Journal of Electrical and Computer Engineering (IJECE)*. *13*(6), 7137–7146. https://doi.org/10.11591/ijece.v13i6.pp7137-7146

Zamroni, M. R., Hamid, M. R., Mujilahwati, S., Sholihin, M., Leksana, D. M. (2025). Detection of Bullying content in online news using a combination of RoBERTA-BiLSTM. *Jurnal Teknik Informatika (JUTIF)*. *6*(1), 41-50. https://doi.org/10.52436/1.jutif.2025.6.1.4140