Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893

Perbandingan *TF-IDF* dan *Sentence-IndoBERT* untuk Klasifikasi Emosi Ulasan Produk Indonesia

DOI: http://dx.doi.org/10.35889/jutisi.v15i4.3872

Miko Kastomo Putro[1*], Mursyid Ardiansyah[2]
[1]Teknik Komputer, Universitas Amikom Yogyakarta, Sleman, Indonesia
[2]Ilmu Komputer, Institut Teknologi Sains dan Bisnis Muhammadiyah Selayar, Selayar, Indonesia
**e-mail Corresponding Author*: miko.putro@amikom.ac.id

Abstract

Informal Indonesian product reviews containing lexical variation posed challenges for emotion classification. This study aimed to compare term frequency-inverse document frequency and Sentence-IndoBERT representations using the same logistic regression classifier so that the effect of text representation could be examined under controlled conditions. The PRDECT-ID dataset was audited, cleaned of duplicates and label conflicts, and reduced to 5,261 reviews. The data were stratified into 70% training, 15% validation, and 15% testing sets. Performance was evaluated using accuracy, macro precision, macro recall, macro F1-score, weighted F1-score, and a confusion matrix. The results showed that term frequency-inverse document frequency achieved 64.05% accuracy and a 60.88% macro F1-score, outperforming Sentence-IndoBERT, which achieved 63.67% accuracy and a 59.28% macro F1-score. The novelty lay in a controlled experiment using the same classifier, data partition, and evaluation procedure. The findings indicated that word-based representation was more effective and computationally efficient for the evaluated dataset.

Keywords: *Emotion Classification; Indonesian Product Reviews; Text Representation; Logistic Regression; Machine Learning*

Abstrak

Ulasan produk berbahasa Indonesia yang singkat, informal, dan mengandung variasi kata telah menimbulkan kesulitan dalam klasifikasi emosi. Penelitian ini telah membandingkan representasi *term frequency-inverse document frequency* dan *Sentence-IndoBERT* menggunakan regresi logistik yang sama agar pengaruh representasi teks dapat diamati secara terkontrol. Dataset PRDECT-ID telah diaudit, dibersihkan dari duplikasi dan konflik label, lalu menghasilkan 5.261 ulasan yang dibagi secara terstratifikasi menjadi 70% data pelatihan, 15% data validasi, dan 15% data pengujian. Kinerja telah dievaluasi menggunakan akurasi, presisi makro, recall makro, skor F1 makro, skor F1 berbobot, dan *confusion matrix*. Hasil pengujian telah menunjukkan bahwa *term frequency-inverse document frequency* memperoleh akurasi 64,05% dan skor F1 makro 60,88%, lebih tinggi daripada *Sentence-IndoBERT* dengan akurasi 63,67% dan skor F1 makro 59,28%. Kebaruan penelitian telah ditunjukkan melalui eksperimen terkontrol dengan pengklasifikasi, pembagian data, dan prosedur evaluasi yang sama. Temuan tersebut telah memperlihatkan bahwa representasi berbasis kata lebih efektif dan efisien untuk dataset yang digunakan.

Kata kunci: *Klasifikasi Emosi; Ulasan Produk Indonesia; Representasi Teks; Logistic Regression; Machine Learning*

1. Pendahuluan

Perkembangan perdagangan elektronik telah menjadikan ulasan pelanggan sebagai sumber informasi penting untuk memahami pengalaman, kepuasan, keluhan, dan respons emosional konsumen terhadap produk maupun layanan. Ulasan tidak hanya memuat penilaian positif atau negatif, tetapi juga dapat mencerminkan emosi yang lebih spesifik, seperti bahagia,

cinta, marah, takut, dan sedih. Dibandingkan analisis sentimen yang umumnya mengelompokkan teks berdasarkan polaritas, klasifikasi emosi mampu memberikan informasi yang lebih terperinci mengenai keadaan emosional konsumen. Informasi tersebut dapat dimanfaatkan untuk mengidentifikasi sumber kepuasan dan ketidakpuasan, mengevaluasi kualitas produk, memperbaiki pelayanan, memprioritaskan penanganan keluhan, serta menyusun strategi respons pelanggan yang lebih tepat. Kebutuhan terhadap klasifikasi emosi semakin penting karena jumlah ulasan produk berbahasa Indonesia terus bertambah dan tidak memungkinkan seluruhnya dianalisis secara manual. *PRDECT-ID* menyediakan 5.400 ulasan produk berbahasa Indonesia dari 29 kategori dengan lima kelas emosi, yaitu *Anger, Fear, Happy, Love, dan Sadness*. Dataset tersebut telah melalui proses anotasi dan validasi sehingga dapat digunakan sebagai sumber data dalam pengembangan dan evaluasi metode klasifikasi emosi pada domain perdagangan elektronik Indonesia [1], [2].

Objek kajian dalam penelitian ini memiliki sejumlah karakteristik yang dapat mempersulit proses klasifikasi. Ulasan produk berbahasa Indonesia umumnya berupa teks pendek dan informal serta mengandung kata tidak baku, singkatan, kesalahan ejaan, istilah asing, negasi, pengulangan huruf, emoji, kapitalisasi yang tidak konsisten, dan ekspresi emosi yang saling berdekatan. Satu kata yang sama juga dapat menghasilkan makna berbeda bergantung pada urutan kata dan konteks kalimat. Sebagai contoh, frasa “barang bagus” dan “barang tidak bagus” memiliki sebagian besar unsur kata yang sama, tetapi menyampaikan makna yang bertentangan. Selain permasalahan kebahasaan, dataset dapat memuat ulasan berulang atau teks identik dengan label emosi yang berbeda. Apabila kondisi tersebut tidak ditangani sebelum pembagian data, ulasan identik berpotensi masuk ke dalam subset pelatihan dan pengujian sehingga menimbulkan kebocoran data dan menghasilkan evaluasi yang terlalu optimistis. Distribusi kelas yang tidak seimbang juga dapat menyebabkan model lebih banyak memprediksi kelas dominan. Oleh karena itu, akurasi saja belum cukup untuk menggambarkan kemampuan model dalam mengenali seluruh kelas emosi. Permasalahan tersebut menunjukkan bahwa klasifikasi emosi pada ulasan produk membutuhkan audit dan pembersihan dataset, pencegahan kebocoran data, pembagian data secara terstratifikasi, pemilihan representasi teks yang sesuai, serta evaluasi yang mempertimbangkan kinerja setiap kelas.

Berbagai penelitian telah berupaya menyelesaikan permasalahan klasifikasi ulasan berbahasa Indonesia melalui representasi statistik maupun representasi kontekstual. Pendekatan berbasis *TF-IDF* telah digunakan untuk mengubah ulasan menjadi vektor berdasarkan tingkat kepentingan kata, kemudian dipadukan dengan pengklasifikasi seperti *Logistic Regression, Random Forest, Naïve Bayes*, dan *Support Vector Machine* [3], [4]. Pendekatan berbasis IndoBERT juga telah diterapkan pada klasifikasi ulasan produk dan ulasan pengguna platform perdagangan elektronik untuk menangkap hubungan semantik serta konteks antarkata [5], [6], [7]. Pada klasifikasi emosi, model *BERT*, *RoBERTa*, dan *DistilBERT* telah digunakan melalui proses fine-tuning untuk mempelajari pola emosi pada teks berbahasa Indonesia [8], [9], [10]. Penelitian lain menunjukkan bahwa distribusi kelas yang tidak seimbang dapat menghasilkan kinerja agregat yang terlihat baik, tetapi model tetap mengalami kesulitan dalam mengenali kelas dengan jumlah data lebih sedikit. Kondisi tersebut memperlihatkan pentingnya penggunaan F1 makro, metrik setiap kelas, dan *confusion matrix* sebagai pelengkap akurasi [11]. Representasi kontekstual dan kombinasi fitur statistik-semantik juga telah diterapkan pada dataset ulasan produk berbahasa Indonesia [2], [12]. Meskipun demikian, penelitian terdahulu umumnya berfokus pada klasifikasi sentimen, membandingkan arsitektur dari keluarga Transformer, melakukan fine-tuning, menggunakan augmentasi data, atau memasangkan setiap representasi dengan pengklasifikasi yang berbeda. Perubahan beberapa komponen secara bersamaan menyebabkan peningkatan atau penurunan kinerja belum dapat dikaitkan secara khusus dengan jenis representasi teks. Dengan demikian, masih terdapat kesenjangan berupa belum tersedianya perbandingan *TF-IDF* dan *Sentence-IndoBERT* untuk klasifikasi lima emosi pada ulasan produk berbahasa Indonesia dengan menggunakan pengklasifikasi, pembagian data, proses validasi, data pengujian, dan prosedur evaluasi yang sama setelah duplikasi serta konflik label ditangani.

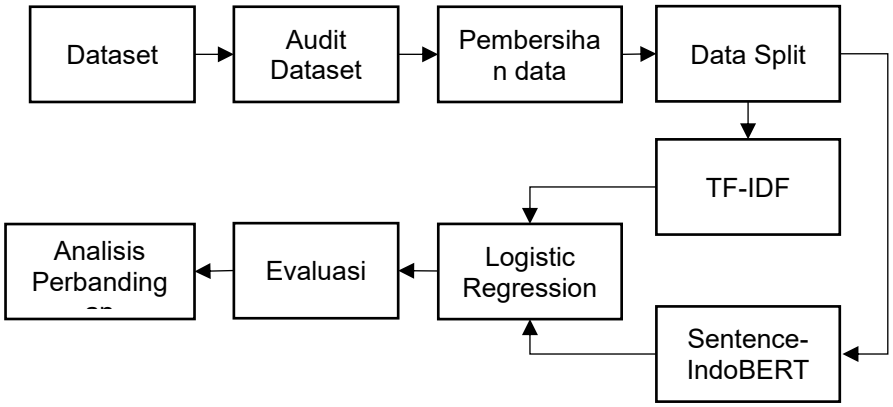
Untuk menjawab kesenjangan tersebut, penelitian ini mengusulkan eksperimen terkontrol yang membandingkan *TF-IDF* sebagai representasi leksikal berbasis kata dengan *Sentence-IndoBERT* sebagai representasi semantik berbasis kalimat. TF-IDF dipilih karena sederhana, ringan, mudah diinterpretasikan, dan mampu memberikan bobot lebih besar pada kata yang dianggap penting dalam dokumen. Penggunaan unigram dan bigram memungkinkan

representasi tersebut menangkap kata tunggal sekaligus pasangan kata yang mengandung intensitas atau negasi, seperti “sangat kecewa”, “tidak sesuai”, dan “bagus sekali”. Pendekatan TF-IDF juga masih relevan untuk klasifikasi ulasan ketika kategori dapat dibedakan melalui kata atau frasa tertentu [3], [4]. Sementara itu, *Sentence-IndoBERT* digunakan karena representasi berbasis Transformer mampu mempertimbangkan urutan kata dan hubungan kontekstual dalam membentuk embedding pada tingkat kalimat [8], [13], [14]. Kedua representasi dipasangkan dengan *Logistic Regression* multikelas yang sama agar pengaruh pengklasifikasi dapat dikendalikan dan perbedaan hasil lebih mencerminkan karakteristik representasi teks [15]. Dataset terlebih dahulu diaudit dan dibersihkan dari duplikasi serta konflik label, kemudian dibagi secara terstratifikasi menjadi data pelatihan, validasi, dan pengujian. State of the art dan kebaruan penelitian ini terletak pada rancangan eksperimen terkontrol yang menjadikan representasi teks sebagai variabel utama, sedangkan pengklasifikasi, pembagian data, proses validasi, data pengujian, dan prosedur evaluasi dipertahankan sama. Kebaruan juga mencakup audit dataset sebelum pembagian data, penghapusan teks duplikat dan konflik label untuk mencegah kebocoran data, serta evaluasi menggunakan akurasi, presisi makro, recall makro, F1 makro, F1 berbobot, metrik setiap kelas, dan *confusion matrix*. Penelitian ini bertujuan membandingkan kinerja *TF-IDF* dan *Sentence-IndoBERT* dalam mengklasifikasikan lima emosi, mengidentifikasi kelas emosi yang paling sulit dikenali, menganalisis pola kesalahan antarkelas, dan memberikan dasar empiris dalam memilih representasi teks yang sesuai untuk klasifikasi emosi pada ulasan produk berbahasa Indonesia.

2. Metodologi

2.1. Rancangan Penelitian

Penelitian ini menggunakan pendekatan eksperimen komputasional terkontrol untuk membandingkan TF-IDF sebagai representasi leksikal berbasis kata dan Sentence-IndoBERT sebagai representasi semantik berbasis kalimat. Kedua representasi dipasangkan dengan Logistic Regression multikelas yang sama. Penggunaan pengklasifikasi, pembagian data, proses validasi, data pengujian, dan prosedur evaluasi yang identik bertujuan agar perbedaan kinerja yang dihasilkan lebih mencerminkan pengaruh representasi teks, bukan perbedaan arsitektur pengklasifikasi atau prosedur pelatihan.



Gambar 1. Alur Penelitian

Prosedur penelitian mengikuti tahapan pada Gambar 1, yaitu: (1) pengumpulan dan penentuan variabel dataset; (2) audit dataset; (3) pembersihan data; (4) pembagian data menjadi data pelatihan, validasi, dan pengujian; (5) pembentukan representasi melalui jalur *TF-IDF* dan *Sentence-IndoBERT*; (6) pelatihan *Logistic Regression*; (7) validasi konfigurasi; (8) evaluasi model; dan (9) analisis perbandingan. *TF-IDF* digunakan untuk membentuk representasi berdasarkan frekuensi dan tingkat kepentingan term, sedangkan *Sentence-IndoBERT* digunakan untuk menghasilkan representasi yang mempertimbangkan hubungan semantik pada tingkat kalimat [7].

2.2. Dataset

Objek pengujian menggunakan *PRDECT-ID* Dataset yang pada kondisi awal terdiri atas 5.400 ulasan produk berbahasa Indonesia. Dataset tersebut memiliki 11 atribut dan mencakup 29 kategori produk dengan lima kelas emosi, yaitu Anger, Fear, Happy, Love, dan Sadness.

Penelitian hanya menggunakan kolom Customer Review sebagai variabel input dan kolom Emotion sebagai variabel target. Kolom lainnya digunakan sebagai metadata pada proses audit atau dikeluarkan dari fitur model untuk mencegah terjadinya kebocoran target.

Tabel 1. Variabel input

Jenis variabel	Kolom	Keterangan
Input	Customer Review	Teks ulasan yang diproses model
Target	Emotion	Kelas emosi hasil klasifikasi
Metadata audit	Category, Product Name, Location	Digunakan untuk memahami kondisi data
Tidak digunakan sebagai fitur	Sentiment, rating, harga, jumlah penjualan	Dikeluarkan agar tidak terjadi kebocoran target

Berdasarkan Tabel 1, *Customer Review* digunakan sebagai sumber informasi utama yang dipelajari oleh model, sedangkan Emotion menjadi label kelas yang harus diprediksi. Kolom *Category, Product Name*, dan *Location* hanya digunakan untuk menganalisis karakteristik dataset selama proses audit. Sementara itu, atribut Sentiment, rating, harga, dan jumlah penjualan tidak dimasukkan sebagai fitur karena dapat memberikan informasi tambahan yang berkaitan dengan kelas emosi. Pembatasan tersebut dilakukan agar model mempelajari pola emosi berdasarkan isi teks ulasan dan agar hasil evaluasi tidak menjadi terlalu optimistis akibat kebocoran target.

Klasifikasi emosi pada penelitian ini merupakan permasalahan multikelas karena setiap ulasan ditempatkan pada satu dari lima kelas emosi. Pengenalan emosi memerlukan pemisahan kelas yang lebih terperinci dibandingkan klasifikasi sentimen biner karena beberapa emosi dapat memiliki polaritas yang sama, tetapi menunjukkan keadaan psikologis yang berbeda [14].

2.3. Audit Dataset

Audit dataset dilakukan sebelum pembersihan dan pembagian data. Tahap ini bertujuan mengetahui kualitas awal dataset serta mengidentifikasi kondisi yang dapat memengaruhi validitas eksperimen. Komponen pemeriksaan audit dataset disajikan pada Tabel berikut.

Tabel 2. Komponen Audit Dataset.

Komponen yang diperiksa	**Tujuan pemeriksaan**
Jumlah baris dan kolom	Mengetahui ukuran awal dataset serta jumlah atribut yang tersedia.
Tipe data setiap atribut	Memastikan setiap atribut memiliki tipe data yang sesuai untuk proses analisis dan pemodelan.
Nilai kosong	Mengidentifikasi data yang tidak memiliki nilai pada atribut tertentu.
Distribusi kelas emosi	Mengetahui jumlah dan proporsi data pada kelas *Anger*, *Fear*, *Happy*, *Love*, dan *Sadness*.
Jumlah kategori produk	Mengetahui keragaman kategori produk yang terdapat dalam dataset.
Duplikasi baris secara penuh	Mengidentifikasi baris yang memiliki nilai identik pada seluruh atribut.
Duplikasi berdasarkan teks ulasan	Mengidentifikasi ulasan dengan isi teks yang sama meskipun atribut lainnya dapat berbeda.
Teks identik dengan label berbeda	Menemukan ulasan yang memiliki teks sama tetapi diberi lebih dari satu label emosi.
Panjang teks berdasarkan jumlah kata	Mengetahui karakteristik panjang ulasan, seperti nilai rata-rata, median, minimum, dan maksimum.
Ulasan dengan format yang tidak dapat diproses	Mengidentifikasi teks rusak, kosong setelah normalisasi, atau memiliki format yang tidak sesuai untuk proses prapemrosesan.

Berdasarkan Tabel 2, audit dataset mencakup pemeriksaan struktur data, kelengkapan atribut, distribusi kelas, duplikasi, konsistensi label, karakteristik panjang teks, dan validitas format ulasan. Hasil audit digunakan sebagai dasar untuk menentukan prosedur pembersihan sebelum dataset dibagi menjadi data pelatihan, validasi, dan pengujian.

Pemeriksaan duplikasi tidak hanya dilakukan terhadap teks asli karena perbedaan huruf kapital, spasi, URL, atau format Unicode dapat menyebabkan dua ulasan yang sebenarnya identik terlihat berbeda. Oleh karena itu, dibentuk teks kunci hasil normalisasi untuk keperluan audit.

Untuk setiap teks asli x_i, normalisasi audit dapat dinyatakan sebagai:

$$ui = S(H\left(U\left(L(NFKC(xi))\right)\right)) \quad (1)$$

Dengan:
$NFKC(\cdot)$ merupakan normalisasi karakter Unicode menggunakan bentuk *Normalization Form Compatibility Composition*;
$L(\cdot)$ merupakan perubahan seluruh huruf menjadi huruf kecil;
$U(\cdot)$ merupakan penghapusan URL;
$H(\cdot)$ merupakan penghapusan tag HTML; dan
$S(\cdot)$ merupakan normalisasi spasi, termasuk penghapusan spasi pada awal dan akhir teks serta penggabungan beberapa spasi menjadi satu spasi.

Teks hasil normalisasi $u_i$hanya digunakan sebagai kunci untuk mendeteksi ulasan identik. Teks tersebut belum menjadi representasi akhir yang dimasukkan ke model karena *TF-IDF* dan *Sentence-IndoBERT* menggunakan prosedur prapemrosesan yang berbeda.

2.4. Pembersihan Dataset

Pembersihan data dilakukan berdasarkan hasil audit. Tahap ini mencakup penanganan nilai kosong, konflik label, duplikasi teks, *URL*, *HTML*, karakter *Unicode*, dan spasi berlebih. Prapemrosesan diperlukan karena ulasan informal dapat mengandung kata tidak baku, kesalahan ejaan, slang, dan variasi bentuk kata yang memengaruhi pembentukan fitur.

Setiap ulasan dikelompokkan berdasarkan teks hasil normalisasi u_i. Kelompok ulasan untuk suatu teks unik u dirumuskan sebagai:

$$G(u) = \{i|ui = u\} \quad (2)$$

Selanjutnya, himpunan label pada kelompok tersebut dinyatakan sebagai:

$$Y(u) = \{yi|i \in G(u)\} \quad (3)$$

Penanganan kelompok duplikasi dan konflik label dilakukan menggunakan aturan:

$$D'^{(u)} = \begin{cases} \{satu\ data\ dari\ G(u)\}, |Y(u)| = 1 \\ \emptyset, |Y(u)| > 1 \end{cases} \quad (4)$$

Apabila seluruh ulasan dalam satu kelompok memiliki label yang sama, satu ulasan dipertahankan dan ulasan lainnya dihapus sebagai duplikasi. Apabila satu teks identik memiliki lebih dari satu label emosi, seluruh data pada kelompok tersebut dihapus karena label dianggap tidak konsisten. Pendekatan penghapusan seluruh kelompok konflik digunakan agar penelitian tidak memilih salah satu label secara subjektif tanpa melakukan anotasi ulang.

Pembersihan dilakukan sebelum pembagian data. Urutan ini penting untuk mencegah satu ulasan muncul pada data pelatihan dan salinannya muncul pada data pengujian. Kondisi tersebut dapat menyebabkan model seolah-olah memiliki kemampuan generalisasi tinggi, padahal model telah menemukan teks yang sama pada tahap pelatihan.

Setelah proses pembersihan selesai, dataset diperiksa kembali untuk memastikan tidak terdapat nilai kosong pada variabel input dan target, tidak terdapat duplikasi berdasarkan teks hasil normalisasi, tidak terdapat teks identik dengan label berbeda, dan setiap data memiliki salah satu dari lima kelas emosi yang valid.

2.5. Pembagian Dataset

Dataset bersih dibagi menggunakan stratified random sampling berdasarkan kelas Emotion agar proporsi setiap kelas tetap terjaga pada data pelatihan, validasi, dan pengujian.

Komposisi pembagian ditetapkan sebesar 70% untuk data pelatihan, 15% untuk data validasi, dan 15% untuk data pengujian.

Tabel 3. Distribusi data training, validation, dan testing.

Kelas emosi	Data bersih	Training	Validation	Testing
Anger	664	465	100	99
Fear	890	623	133	134
Happy	1.746	1.222	262	262
Love	793	555	119	119
Sadness	1.168	817	175	176
Total	5.261	3.682	789	790

Berdasarkan Tabel 3, sebanyak 5.261 data bersih dibagi menjadi 3.682 data pelatihan, 789 data validasi, dan 790 data pengujian. Kelas *Happy* memiliki jumlah data terbesar pada seluruh subset, sedangkan kelas *Anger* memiliki jumlah paling sedikit. Meskipun jumlah data setiap kelas berbeda, pembagian terstratifikasi mempertahankan proporsi setiap kelas secara relatif konsisten.

2.6. Representasi TF-IDF

Jalur pertama menggunakan *TF-IDF* untuk mengubah setiap ulasan menjadi vektor numerik. Pada jalur ini, teks melalui tahapan prapemrosesan berupa normalisasi Unicode, perubahan huruf menjadi kecil, penghapusan *URL* dan *HTML*, penghapusan tanda baca yang tidak diperlukan, normalisasi spasi, serta normalisasi kata tidak baku dan istilah asing menggunakan kamus normalisasi. Kata negasi seperti belum, bukan, jangan, kurang, mustahil, tanpa, dan tidak tetap dipertahankan karena dapat mengubah makna suatu frasa; sebagai contoh, kata “bagus” dapat menunjukkan penilaian positif, sedangkan frasa “tidak bagus” menyampaikan penilaian yang berbeda. *Stemming* tidak diterapkan pada konfigurasi akhir agar bentuk kata dan frasa emosional tidak kehilangan informasi yang dapat membedakan kelas. Setelah proses tersebut, teks ditokenisasi menjadi unigram dan bigram, dengan unigram merepresentasikan satu kata dan bigram merepresentasikan dua kata yang muncul secara berurutan. Penggunaan kedua jenis fitur ini memungkinkan model mempelajari kata maupun frasa seperti “kecewa”, “bagus”, “tidak sesuai”, “sangat kecewa”, dan “bagus sekali”.

Apabila *f(t,d)* merupakan jumlah kemunculan term *t* dalam dokumen *d*, nilai term frequency dirumuskan sebagai:

$$TF(t,d) = \frac{f(t,d)}{\sum_{t^i \in d} f(t^i d)} \tag{5}$$

Apabila N merupakan jumlah dokumen dalam data pelatihan dan $df(t)$ merupakan jumlah dokumen yang memuat term t, nilai *inverse document frequency* dirumuskan sebagai:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \tag{6}$$

Bobot TF-IDF untuk term $t$pada dokumen $d$dihitung menggunakan:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \tag{7}$$

TF-IDF hanya di-fit menggunakan data pelatihan. Kosakata dan nilai *IDF* yang diperoleh dari data pelatihan kemudian digunakan untuk mentransformasikan data validasi dan pengujian. Prosedur ini dilakukan agar informasi dari data validasi dan pengujian tidak memengaruhi pembentukan fitur.

2.7. Representasi Sentence-IndoBERT

Jalur kedua menggunakan *Sentence-IndoBERT* untuk membentuk *embedding* padat pada tingkat kalimat. *IndoBERT* merupakan model bahasa pralatih menggunakan korpus bahasa

Indonesia sehingga dapat merepresentasikan pola kosakata dan konteks bahasa Indonesia dengan lebih sesuai dibandingkan representasi berbasis frekuensi semata [19].

Berbeda dari jalur *TF-IDF*, jalur *Sentence-IndoBE*RT mempertahankan struktur teks yang lebih dekat dengan bentuk aslinya. Huruf, tanda baca, urutan kata, dan konteks kalimat dipertahankan sejauh dapat diproses oleh *tokenizer*. Prapemrosesan dibatasi pada penanganan nilai kosong, normalisasi format Unicode, penghapusan *URL* atau *HTML* yang tidak mengandung informasi emosi, dan normalisasi spasi.

Setiap ulasan ditokenisasi menggunakan *tokenizer* dari model *Sentence-IndoBERT*. Token khusus ditambahkan sesuai kebutuhan model, kemudian urutan token dipotong atau diberi padding hingga panjang maksimum yang ditentukan. Hasil tokenisasi terdiri atas input_ids dan attention_mask.

Untuk suatu kalimat, representasi awal dapat dituliskan sebagai:

$$H = [h_1, h_2, \dots, h_t] \tag{8}$$

dengan:
h_i merupakan hidden state token ke-i;
T merupakan jumlah token; dan
H merupakan keluaran lapisan Transformer.

Embedding kalimat dibentuk menggunakan *mean pooling* dengan memperhatikan attention_mask:

$$s = \frac{\sum_{i=1}^{T} m_i h_i}{\sum_{i=1}^{T} m_i} \tag{9}$$

dengan:
s merupakan embedding kalimat;
h_i merupakan embedding token ke-i;
$m_i \in \{0, 1\}$ merupakan nilai attention_mask; dan
T merupakan panjang urutan token.

Nilai $m_i = 1$ menunjukkan *token* asli yang diperhitungkan dalam proses *pooling*, sedangkan $m_i = 0$menunjukkan *token padding* yang tidak diperhitungkan. Dengan demikian, token *padding* tidak memengaruhi *embedding* kalimat.

Sentence-IndoBERT digunakan dalam kondisi *frozen*. Artinya, bobot model *Transformer* tidak diperbarui selama pelatihan Logistic Regression. Setiap ulasan hanya dilewatkan melalui model untuk memperoleh embedding tetap. Pendekatan ini digunakan agar perbandingan berfokus pada kualitas representasi yang dihasilkan *TF-IDF* dan *Sentence-IndoBERT*, tanpa mencampurkan pengaruh proses *fine-tuning*.

2.8. Logistic Regression

Vektor *TF-IDF* dan embedding *Sentence-IndoBERT* digunakan secara terpisah sebagai masukan bagi *Logistic Regressio*n multikelas. Pengklasifikasi yang sama diterapkan pada kedua jenis representasi agar perbedaan hasil klasifikasi lebih mencerminkan pengaruh representasi teks dan tidak dipengaruhi oleh penggunaan pengklasifikasi yang berbeda. Model *Logistic Regression* dilatih menggunakan data pelatihan untuk mempelajari hubungan antara fitur masukan dan lima kelas emosi, yaitu *Anger, Fear, Happy, Love*, dan *Sadness*.

Untuk setiap vektor masukan x, nilai logit kelas ke-k dihitung menggunakan persamaan berikut:

$$z_k = \beta_k^T x + b_k \tag{10}$$

dengan:
x merupakan vektor *TF-IDF* atau embedding *Sentence-IndoBERT*;
β_k merupakan vektor koefisien untuk kelas ke-k;
b_k merupakan bias kelas ke-k; dan
$k \in \{1, 2, \dots, 5\}$.

Probabilitas setiap kelas dihitung menggunakan fungsi *softmax*:

$$P(y = k|x) = \frac{\exp\left(\beta_k^T x + b_k\right)}{\sum_{j=1}^{k} \exp\left(\beta_j^T x + b_j\right)} \quad (11)$$

dengan:
$P(y = k \mid x)$ merupakan probabilitas bahwa data dengan vektor fitur x termasuk ke dalam kelas ke-k;
x merupakan vektor fitur *TF-IDF* atau *embedding Sentence-IndoBERT*;
β_k merupakan vektor koefisien *Logistic Regression* untuk kelas ke-k;
b_k merupakan nilai bias untuk kelas ke-k;
j merupakan indeks kelas yang digunakan dalam perhitungan total probabilitas;
k merupakan kelas emosi yang sedang dihitung probabilitasnya;
$K = 5$ merupakan jumlah kelas emosi; dan
$\exp(\cdot)$ merupakan fungsi eksponensial.

Fungsi *softmax* menghasilkan nilai probabilitas untuk seluruh kelas dengan jumlah probabilitas sama dengan satu. Kelas dengan probabilitas tertinggi ditetapkan sebagai hasil prediksi model. *Logistic Regression* pada kedua jalur dilatih menggunakan data pelatihan, sedangkan pemilihan konfigurasi terbaik dilakukan berdasarkan kinerja pada data validasi. Setelah konfigurasi terbaik diperoleh, model dievaluasi menggunakan data pengujian yang sama agar perbandingan *TF-IDF* dan *Sentence-IndoBERT* dilakukan secara objektif.

2.9. Evaluasi

Validasi konfigurasi dilakukan pada data validation, sedangkan evaluasi akhir dilakukan pada data testing yang sama untuk kedua model. Metrik yang digunakan adalah *accuracy*, *precision*, *recall*, *F1-score*, *Macro Precision*, *Macro Recall*, *Macro F1*, *Weighted F1*, *dan confusion matrix.*

1. *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (12)$$

2. *Precision*

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

3. *Recall*

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

4. *F1 - Score*

$$F1 - Score = 2X \frac{Precision\ X\ Recall}{Precision + Recall} \quad (15)$$

(*TP* atau *True Positive* adalah Logistic Regression memprediksi Positif dan data aktualnya Positif (Benar), *TN* atau *True Negative* adalah Logistic Regression memprediksi Negatif dan data aktualnya Negatif (Benar). *FP* atau *False Positive* adalah Logistic Regression memprediksi Positif dan data aktualnya Negatif (Salah), *FN* atau *False Negative* adalah Logistic Regression memprediksi Negatif dan data aktualnya Positif (Salah).

3. Hasil dan Pembahasan

3.1. Hasil

Prapemrosesan dilakukan terhadap 5.400 ulasan awal dengan mengikuti tahapan pemeriksaan nilai kosong, penanganan konflik label, penghapusan duplikasi, normalisasi teks, serta pembentukan dua jenis teks untuk *TF-IDF* dan *Sentence-IndoBERT*. Tabel 4 menyajikan perubahan jumlah data dan konfigurasi utama yang digunakan selama proses pembersihan dan prapemrosesan.

Tabel 4. Hasil pembersihan dan prapemrosesan dataset.

Tahapan atau komponen	Hasil
Jumlah data awal	5.400
Baris kosong yang dihapus	0
Kelompok teks dengan konflik label	27
Baris yang dihapus karena konflik label	57
Baris dalam kelompok duplikasi berlabel sama	146
Baris duplikat yang dihapus	82
Jumlah data setelah pembersihan	5.261
Duplikasi teks yang tersisa	0
Teks kosong setelah prapemrosesan	0
Jumlah pasangan dalam kamus normalisasi	40
Jumlah kata negasi yang dipertahankan	7
Penerapan *stemming*	Tidak
Jumlah data akhir yang dapat diproses	5.261

Berdasarkan Tabel 4, proses pembersihan mengurangi jumlah data dari 5.400 menjadi 5.261 ulasan. Sebanyak 139 baris atau sekitar 2,57% dari data awal dikeluarkan. Pengurangan tersebut terdiri atas 57 baris yang berada dalam kelompok konflik label dan 82 baris duplikat dengan label yang sama. Setelah pembersihan, tidak ditemukan lagi duplikasi teks maupun teks kosong sehingga seluruh 5.261 ulasan dapat digunakan pada tahap berikutnya. Prapemrosesan juga menggunakan kamus normalisasi yang terdiri atas 40 pasangan kata dan mempertahankan tujuh bentuk negasi, yaitu belum, bukan, jangan, kurang, mustahil, tanpa, dan tidak.

Prapemrosesan menghasilkan dua bentuk teks yang berbeda. Teks untuk *TF-IDF* dinormalisasi agar variasi kosakata berkurang, sedangkan teks untuk *Sentence-IndoBERT* dipertahankan mendekati bentuk aslinya. Contoh hasil kedua jalur prapemrosesan disajikan pada Tabel 5.

Tabel 5. Contoh hasil prapemrosesan TF-IDF dan Sentence-IndoBERT.

Emosi	Teks asli	Teks TF-IDF	Teks Sentence-IndoBERT
Happy	barang bagus dan respon cepat, harga bersaing dengan yg lain.	barang bagus dan respons cepat harga bersaing dengan yang lain	barang bagus dan respon cepat, harga bersaing dengan yg lain.
Sadness	Mic dari awal ga nyala	mic dari awal tidak nyala	Mic dari awal ga nyala
Anger	BARANG KW JANGAN BELI DISINI	barang kw jangan beli disini	BARANG KW JANGAN BELI DISINI
Love	Recommended Seller!! Respon cepat, pengiriman cukup cepat...	recommended penjual respons cepat pengiriman cukup cepat	Recommended Seller!! Respon cepat, pengiriman cukup cepat..
Fear	Gak bisa di gunakan ini gimana	tidak bisa di gunakan ini gimana	Gak bisa di gunakan ini gimana

Berdasarkan Tabel 5, jalur *TF-IDF* mengubah teks menjadi huruf kecil, menghapus tanda baca yang tidak diperlukan, dan menormalisasi beberapa bentuk kata. Kata respon dinormalisasi menjadi respons, yg menjadi yang, ga dan gak menjadi tidak, serta seller menjadi penjual. Sebaliknya, jalur *Sentence-IndoBERT* mempertahankan kapitalisasi, tanda baca, kata tidak baku, dan struktur kalimat yang mendekati teks asli. Perbedaan tersebut menghasilkan vektor *TF-IDF* berbasis unigram dan bigram serta embedding *Sentence-IndoBERT* berdimensi 768.

Setelah prapemrosesan selesai, 5.261 ulasan dibagi secara terstratifikasi menjadi data pelatihan, validasi, dan pengujian. Tabel 6 menyajikan jumlah data setiap kelas pada ketiga subset tersebut.

Tabel 6. Distribusi data training, validation, dan testing

Kelas emosi	Data bersih	Training	Validation	Testing
Anger	664	465	100	99
Fear	890	623	133	134
Happy	1.746	1.222	262	262
Love	793	555	119	119
Sadness	1.168	817	175	176
Total	5.261	3.682	789	790

Berdasarkan Tabel 6, data pelatihan terdiri atas 3.682 ulasan, data validasi terdiri atas 789 ulasan, dan data pengujian terdiri atas 790 ulasan. Proporsi setiap kelas relatif konsisten pada seluruh subset. Kelas *Happy* memiliki jumlah data terbesar dengan proporsi sekitar 33%, sedangkan kelas *Anger* memiliki jumlah data paling sedikit dengan proporsi sekitar 12,6%. Pemeriksaan setelah pembagian menunjukkan bahwa tidak terdapat teks identik antara data pelatihan dan validasi, data pelatihan dan pengujian, maupun data validasi dan pengujian.

Evaluasi akhir dilakukan menggunakan 790 data pengujian yang sama untuk kedua model. Perbandingan dilakukan berdasarkan jumlah prediksi benar, akurasi, *presisi* makro, *recall* makro, F1 makro, dan F1 berbobot. Hasil pengujian disajikan pada Tabel 7.

Tabel 7. Perbandingan kinerja model pada data pengujian

Model	Prediksi benar	Akurasi (%)	Presisi makro (%)	Recall makro (%)	F1 makro (%)	F1 berbobot (%)
TF-IDF + Logistic Regression	506	64,05	60,99	61,12	60,88	64,71
Sentence-IndoBERT + Logistic Regression	503	63,67	60,04	58,79	59,28	63,25

Berdasarkan Tabel 7, *TF-IDF* menghasilkan 506 prediksi benar, sedangkan *Sentence-IndoBERT* menghasilkan 503 prediksi benar. *TF-IDF* memperoleh nilai lebih tinggi pada seluruh metrik agregat. Selisih akurasi kedua model relatif kecil, yaitu sekitar 0,38 poin persentase. Selisih terbesar terdapat pada recall makro sebesar 2,33 poin persentase, diikuti F1 makro sekitar 1,61 poin persentase dan F1 berbobot sekitar 1,45 poin persentase.

Selain metrik agregat, evaluasi dilakukan terhadap presisi, recall, dan F1-score pada masing-masing kelas emosi. Tabel 8 menyajikan hasil *TF-IDF* untuk kelas *Anger, Fear, Happy, Love,* dan *Sadness*.

Tabel 8. Hasil TF-IDF pada setiap kelas emosi

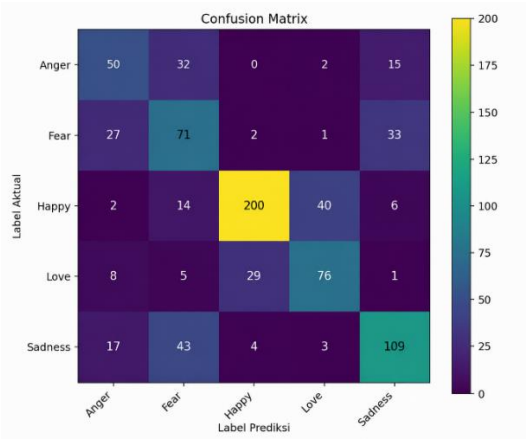
Emosi	Presisi (%)	Recall (%)	F1-score (%)	Jumlah data
Anger	48,08	50,51	49,26	99
Fear	43,03	52,99	47,49	134
Happy	85,11	76,34	80,48	262
Love	62,30	63,87	63,07	119
Sadness	66,46	61,93	64,12	176

Berdasarkan Tabel 8, *TF-IDF* memperoleh kinerja tertinggi pada kelas *Happy*, dengan presisi 85,11%, *recall* 76,34%, dan *F1-score* 80,48%. Kelas *Sadness* dan *Love* masing-masing menghasilkan *F1-score* 64,12% dan 63,07%. *F1-score* terendah terdapat pada kelas *Fear* sebesar 47,49%, diikuti kelas *Anger* sebesar 49,26%. Hasil pengujian setiap kelas pada *Sentence-IndoBERT* dapat dilihat pada Tabel 9.

Tabel 9. Hasil Sentence-IndoBERT pada setiap kelas emosi

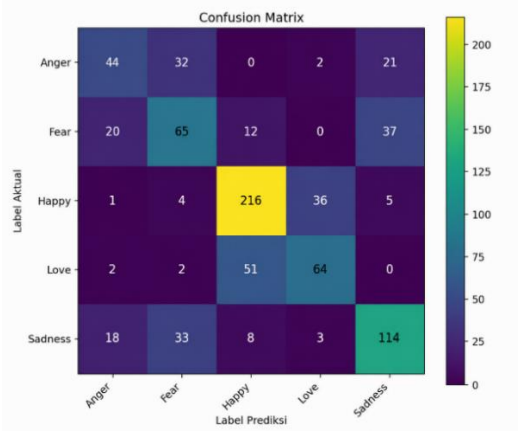
Emosi	Presisi (%)	Recall (%)	F1-score (%)	Jumlah data
Anger	51,76	44,44	47,83	99
Fear	47,79	48,51	48,15	134
Happy	75,26	82,44	78,69	262
Love	60,95	53,78	57,14	119
Sadness	64,41	64,77	64,59	176

Berdasarkan Tabel 9, *Sentence-IndoBERT* memperoleh *F1-score* tertinggi pada kelas *Happy* sebesar 78,69%, diikuti *Sadness* sebesar 64,59%. Kinerja terendah terdapat pada kelas *Anger* dengan *F1-score* 47,83% dan kelas Fear dengan *F1-score* 48,15%. *Sentence-IndoBERT* memperoleh recall *Happy* sebesar 82,44%, tetapi presisinya sebesar 75,26%. Pada kelas *Love*, *Sentence-IndoBERT* menghasilkan *F1-score* 57,14%. Pola prediksi benar dan kesalahan antarkelas pada *TF-IDF* divisualisasikan menggunakan *confusion matrix* pada Gambar 2



Gambar 2. Confusion Matrix TF-IDF

Berdasarkan Gambar 2, model *TF-IDF* mampu mengklasifikasikan dengan benar 50 data '*Anger*', 71 data '*Fear*', 200 data '*Happy*', 76 data '*Love*', dan 109 data '*Sadness*'. Kelas '*Happy*' menunjukkan jumlah prediksi benar tertinggi, sedangkan kesalahan terbesar terjadi ketika 40 data '*Happy*' diprediksi sebagai '*Love*'. Pada kelas negatif, 32 data '*Anger*' salah diprediksi sebagai '*Fear*', sedangkan 33 data '*Fear*' diprediksi sebagai 'Sadness'. Kelas '*Sadness*' juga cukup sering tertukar dengan '*Fear*', yaitu sebanyak 43 data. Sementara itu, kelas '*Love*' paling banyak salah diprediksi sebagai '*Happy*', yaitu 29 data. Pola ini menunjukkan bahwa *TF-IDF* cukup baik mengenali kelas '*Happy*', '*Love*', dan '*Sadness*', tetapi masih mengalami kesulitan membedakan kelas yang memiliki kemiripan kosakata atau ekspresi, terutama '*Anger*' dengan '*Fear*', '*Fear*' dengan '*Sadness*', serta '*Happy*' dengan '*Love*'. Pola prediksi *Sentence-IndoBERT* dapat dilihat melalui confusion matrix pada Gambar 3.



Gambar 3. Confusion Matrix Sentence Indo-BERT

Berdasarkan Gambar 3, model *Sentence-IndoBERT* mampu mengklasifikasikan dengan benar 44 data '*Anger*', 65 data '*Fear*', 216 data '*Happy*', 64 data '*Love*', dan 114 data '*Sadness*'. Kelas '*Happy*' memperoleh jumlah prediksi benar tertinggi, tetapi masih terdapat 36 data '*Happy*' yang diprediksi sebagai '*Love*'. Kesalahan paling menonjol pada kelas '*Love*' terjadi ketika 51 data diprediksi sebagai '*Happy*', yang menunjukkan adanya kemiripan konteks antara kedua emosi positif tersebut. Pada kelas negatif, 32 data '*Anger*' diprediksi sebagai '*Fear*' dan 21 data lainnya sebagai '*Sadness*', sedangkan 37 data '*Fear*' salah diprediksi sebagai '*Sadness*'. Kelas '*Sadness*' juga masih tertukar dengan '*Fear*' sebanyak 33 data dan dengan '*Anger*' sebanyak 18 data. Pola ini menunjukkan bahwa Sentence-IndoBERT cukup baik mengenali kelas '*Happy*' dan '*Sadness*', tetapi masih mengalami kesulitan dalam membedakan emosi yang memiliki kedekatan makna, terutama '*Happy*' dengan '*Love*' serta '*Anger*', '*Fear*', dan '*Sadness*'.

3.2. Pembahasan

Hasil penelitian menunjukkan bahwa *TF-IDF* menghasilkan akurasi 64,05%, F1 makro 60,88%, dan F1 berbobot 64,71%, sedikit lebih tinggi dibandingkan *Sentence-IndoBERT* yang memperoleh masing-masing 63,67%, 59,28%, dan 63,25%. Hasil tersebut menjawab tujuan penelitian bahwa, ketika pengklasifikasi, pembagian data, dan prosedur evaluasi disamakan, *TF-IDF* lebih sesuai untuk karakteristik *PRDECT-ID* yang didominasi ulasan pendek dan penanda emosi eksplisit. Penggunaan unigram dan bigram memungkinkan kata serta frasa seperti "kecewa", "tidak sesuai", dan "sangat bagus" direpresentasikan secara langsung. Hal ini sejalan dengan konsep bahwa metode statistik dan pengklasifikasi linear tetap kompetitif ketika kelas dapat dibedakan melalui pola leksikal yang kuat [16]. Sebaliknya, *Sentence-IndoBERT* secara teoritis mampu memahami konteks, tetapi pada penelitian ini digunakan dalam kondisi *frozen* sehingga representasinya tidak disesuaikan secara langsung dengan karakteristik lima kelas emosi. Kinerja model berbasis *BERT* memang dipengaruhi oleh kesesuaian domain, strategi adaptasi, dan konfigurasi pelatihan [17].

Dibandingkan penelitian sebelumnya yang memperoleh peningkatan melalui fine-tuning, perubahan arsitektur, augmentasi, atau penggunaan pengklasifikasi berbeda, penelitian ini memiliki keunggulan metodologis berupa eksperimen yang lebih terkontrol. *TF-IDF* dan *Sentence-IndoBERT* menggunakan *Logistic Regression*, data validasi, data pengujian, serta metrik evaluasi yang sama, sehingga perbedaan hasil lebih dapat dikaitkan dengan representasi teks. Penelitian ini juga mengevaluasi lima kelas emosi melalui akurasi, F1 makro, F1 berbobot, metrik setiap kelas, confusion matrix, dan waktu komputasi. Hasil per kelas menunjukkan bahwa *TF-IDF* lebih baik pada *Anger, Happy,* dan *Love*, sedangkan *Sentence-IndoBERT* sedikit lebih baik pada *Fear* dan *Sadness*. Pola kesalahan antara *Happy–Love dan Anger–Fear–Sadness* memperlihatkan bahwa kedua representasi menangkap informasi yang berbeda. Dibandingkan kajian klasifikasi teks yang umumnya membahas algoritma secara terpisah, penelitian ini memberikan temuan baru mengenai perbandingan leksikal dan kontekstual setelah duplikasi serta konflik label ditangani [18].

Kontribusi penelitian ini bagi klasifikasi emosi bahasa Indonesia terletak pada penyediaan prosedur evaluasi yang transparan, terkontrol, dan lebih rendah risiko kebocoran data. Penghapusan 57 data berkonflik dan 82 duplikat sebelum pembagian data memperkuat validitas hasil. Namun, penelitian masih terbatas pada satu dataset, satu pembagian data, satu pengklasifikasi, dan *Sentence-IndoBERT* tanpa *fine-tuning*. Penggunaan satu pembagian data dapat memengaruhi kestabilan estimasi kinerja, sehingga penelitian selanjutnya perlu menggunakan beberapa *random seed*, validasi silang berulang, dan dataset eksternal [19]. Pengembangan berikutnya juga dapat menguji *fine-tuning*, adaptasi domain, penanganan ketidakseimbangan kelas, serta penggabungan *TF-IDF* dan *Sentence-IndoBERT* untuk memanfaatkan informasi leksikal dan kontekstual secara bersamaan.

4. Simpulan

Penelitian ini menunjukkan bahwa *TF-IDF* dan *Sentence-IndoBERT* dapat digunakan untuk mengklasifikasikan lima emosi pada ulasan produk berbahasa Indonesia, tetapi *TF-IDF* memberikan hasil yang lebih baik ketika kedua representasi dipasangkan dengan *Logistic Regression* serta diuji menggunakan pembagian data dan prosedur evaluasi yang sama. Setelah audit, penghapusan konflik label, dan deduplikasi, diperoleh 5.261 ulasan unik yang dibagi menjadi 3.682 data training, 789 data validation, dan 790 data testing. Pada pengujian akhir, TF-IDF menghasilkan accuracy 64,05%, *Macro Precision* 60,99%, Macro Recall 61,12%, Macro F1

60,88%, dan Weighted F1 64,71%, sedangkan *Sentence-IndoBERT* menghasilkan *accuracy* 63,67%, *Macro Precision* 60,04%, *Macro Recall* 58,79%, *Macro F1* 59,28%, dan *Weighted F1* 63,25%. *TF-IDF* juga lebih unggul pada kelas *Anger, Happy,* dan *Love*, serta memiliki waktu pemrosesan sekitar 66,88 kali lebih cepat dibandingkan *Sentence-IndoBERT* pada lingkungan pengujian. Temuan ini menjawab tujuan penelitian bahwa representasi yang lebih kompleks tidak selalu menghasilkan kinerja lebih tinggi, khususnya pada ulasan produk yang relatif pendek dan banyak mengandung kata maupun frasa emosional eksplisit. Kebaruan penelitian terletak pada eksperimen terkontrol yang mengisolasi pengaruh representasi teks melalui penggunaan classifier, pembagian data, proses validasi, dan data testing yang sama. Berdasarkan hasil tersebut, *TF-IDF* lebih sesuai digunakan ketika sistem mengutamakan efisiensi, kesederhanaan, dan kinerja klasifikasi yang stabil, sedangkan *Sentence-IndoBERT* tetap memiliki potensi untuk dikembangkan melalui *fine-tuning*, penyesuaian domain, penanganan ketidakseimbangan kelas, perluasan kamus normalisasi, serta penggunaan dataset yang lebih besar dan lebih beragam agar kemampuan representasi kontekstualnya dapat dimanfaatkan secara lebih optimal.

Daftar Referensi

[1] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, dan S. M. Isa, “PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks,” *Data Brief*, vol. 44, pp. 108554, Okt 2022, doi: 10.1016/J.DIB.2022.108554.

[2] A. D. P. Ariyanto, Fari Katul Fikriah, dan Arif Fitra Setyawan, “Emotion Detection Using Contextual Embeddings for Indonesian Product Review Texts on E-commerce Platform,” *Pixel :Jurnal Ilmiah Komputer Grafis*, vol. 17, no. 1, pp. 179–185, Jul 2024, doi: 10.51903/pixel.v17i1.2010.

[3] V. B. Lestari dan C. A. Hutagalung, “Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis of Indonesian-Language Marketplaces Using SVM, Logistic Regression, and Naive Bayes,” *J-KOMA Journal of Computer Science and Applications*, vol. 8, no. 1, pp. 22–2025, 2025, doi: 10.21009/j.

[4] R. D. Kurniawan, “GoPay App Review Sentiment Classification Optimization Using a Combination of Text Representation and Machine Learning,” 2024. [Daring]. Tersedia pada: http://ejournal.uksw.edu/ijiteb

[5] S. Aras, M. Yusuf, R. Ruimassa, E. A. B. Wambrauw, dan E. B. Palalangan, “Sentiment Analysis on Shopee Product Reviews Using IndoBERT,” *ISI Journal of Information System and Informatics*, vol. 6, no. 6, pp. 132-141, 2024, Diakses: 25 Juni 2026. [Daring]. Tersedia pada: https://doi.org/10.51519/journalisi.v6i3.814

[6] K. C. Pradhisa dan R. Fajriyah, “Analisis Sentimen Ulasan Pengguna E-commerce di Google Play Store Menggunakan Metode IndoBERT,” *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 1, pp. 68-79, Jun 2024, doi: 10.47065/bits.v6i1.5247.

[7] K. Wau, “Application of Fine-Tuned IndoBERT for Sentiment Classification Local Product Reviews on Tokopedia Marketplace with Limited Dataset,” *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 2808–4519, 2025, [Daring]. Tersedia pada: https://ioinformatic.org/

[8] A. S. Rizky dan E. Y. Hidayat, “Emotion Classification in Indonesian Text Using IndoBERT,” *Computer Engineering and Applications*, vol. 14, no. 1, pp. 2252–4274, 2025.

[9] J. I. T. Krisna, A. Luthfiarta, L. D. Cahya, S. Winarno, dan A. Nugraha, “Comparing Optimizer Strategies For Enhancing Emotion Classification In IndoBERT Models,” *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 2, pp. 12-21, Feb 2024, doi: 10.26877/asset.v6i2.18228.

[10] F. Basbeth dan D. H. Fudholi, “Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan Distil-BERT,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, pp. 1160, Apr 2024, doi: 10.30865/mib.v8i2.7472.

[11] Y. A. Singgalen, “Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data,” *Journal of Information System Research (JOSH)*, vol. 6, no. 2, pp. 976–986, Jan 2025, doi: 10.47065/josh.v6i2.6505.

[12] A. Devi, P. Ariyanto, F. Katul Fikriah, dan A. F. Setyawan, “Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences,” *CommIT (Communication and Information Technology) Journal*, vol. 19, no. 1, pp. 1-13. 2025.

[13] M. Habib Algifari dan D. Nugroho, “Emotion Classification of Indonesian Tweets using BERT Embedding,” *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 172-176. 2023. [Daring]. Tersedia pada: http://jurnal.polibatam.ac.id/index.php/JAIC

[14] H. Hartatik dan A. Syafrianto, “Penerapan Model Sentence-BERT Untuk Sistem Rekomendasi Buku Berbasis Konten Di Perpustakaan Digital,” *Jurnal Dialektika Informatika (Detika)*, vol. 6, no. 1, pp. 12–19, Nov 2025, doi: 10.24176/detika.v6i1.15916.

[15] S. Pradeepa *dkk.*, “HGATT_LR: transforming review text classification with hypergraphs attention layer and logistic regression,” *Sci. Rep.*, vol. 14, no. 1, pp. 69-77, Des 2024, doi: 10.1038/s41598-024-70565-6.

[16] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, dan D. Brown, “Text classification algorithms: A survey,” 2019, *MDPI AG*. doi: 10.3390/info10040150.

[17] M. Khadhraoui, H. Bellaaj, M. Ben Ammar, H. Hamam, dan M. Jmaiel, “Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study,” *Applied Sciences (Switzerland)*, vol. 12, no. 6, pp. 23-31, Mar 2022, doi: 10.3390/app12062891.

[18] A. Gasparetto, M. Marcuzzo, A. Zangari, dan A. Albarelli, “ Survey on Text Classification Algorithms: From Text to Predictions,” *Information (Switzerland)*, vol. 13, no. 2, pp. 112-123, Feb 2022, doi: 10.3390/info13020083.

[19] A. Vabalas, E. Gowen, E. Poliakoff, dan A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLoS One*, vol. 14, no. 11, pp. 281-292, Nov 2019, doi: 10.1371/journal.pone.0224365.