

Ablation Study of Attention Pooling and Adaptive Focal Loss for Imbalanced Indonesian Sentiment Classification Using IndoBERT

Aldi Rifaldi Saputra^{1*}, Eka Miranda^{2**}

* Information Systems Management, Department, BINUS Graduate Program, Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia

** Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
aldi.saputra608@binus.ac.id¹, ekamiranda@binus.ac.id²

Article Info

Article history:

Received 2026-09-18

Revised 2026-10-14

Accepted 2026-10-18

Keyword:

Sentiment analysis,
IndoBERT,
Focal loss,
Supervised Contrastive Loss,
Naive Bayes.

ABSTRACT

Sentiment classification of Indonesian YouTube comments remains challenging because informal language, target ambiguity, and severe class imbalance can obscure minority-class performance. This study evaluates the individual and combined contributions of attention pooling, supervised contrastive loss (SCL), and dynamic adaptive focal loss to IndoBERT using a complete 2×2×2 full-factorial ablation design. From an initial 75,453 comments, temporal filtering using 21 August 2026 as the case reference date retained 67,633 comments. After removing 159 irrelevant comments and 2,721 exact duplicates, 64,753 weak-labelled comments remained, with a majority-to-minority ratio of 22.24. A disproportionate stratified sample of 500 comments was independently audited by a human annotator; among 498 valid sentiment items, LLM-to-human agreement reached a Macro-F1 of 0.5389 and Cohen's kappa of 0.3146, indicating substantial weak-label noise. The dataset was split 80:10:10 using a fixed stratified split, and each IndoBERT variant was trained with three seeds (42, 123, and 3407). Plain IndoBERT achieved a mean Macro-F1 of 0.9566 ± 0.0064 . The highest observed mean was obtained by the dynamic-focal-only variant at 0.9595 ± 0.0023 , closely followed by the full configuration at 0.9595 ± 0.0011 . Paired marginal comparisons showed small, context-dependent effects, with all 95% confidence intervals crossing zero. Overall, most gains over TF-IDF baselines came from IndoBERT's contextual representations, while the added components produced comparatively small and unstable improvements.



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

I. PENDAHULUAN

Kepercayaan merupakan aset utama industri perbankan, dan tekanan terhadap kepercayaan tersebut kini terbentuk di ruang publik digital. Penelitian pada industri penerbangan menunjukkan bahwa krisis merek mengubah pola komunikasi perusahaan di media sosial secara terukur, termasuk menurunnya ketanggapan justru ketika perhatian publik memuncak [1]. Kondisi serupa muncul pada kontroversi pemblokiran rekening seorang nasabah Bank Mandiri pada 21 Agustus 2026. Peristiwa tersebut memicu gelombang komentar berbahasa Indonesia dalam volume besar pada kanal *YouTube*, sehingga tersedia data teks yang dapat dianalisis untuk memetakan respons publik terhadap institusi perbankan.

Komentar *YouTube* berbahasa Indonesia bukan objek yang mudah diklasifikasikan. Teksnya pendek, sarat singkatan, bercampur kode kedaerahan, memuat emoji, dan sering mengandung sarkasme. Karakteristik tersebut telah dilaporkan berulang pada korpus komentar *YouTube* berbahasa Indonesia bertema kebijakan publik [2], [3] dan bertema demonstrasi [4]. Di samping persoalan kebahasaan, distribusi kelas pada korpus semacam ini jarang seimbang. Pada kasus yang diteliti, komentar bermuatan positif terhadap bank hanya menempati porsi yang sangat kecil, sementara komentar netral mendominasi. Kondisi tersebut membuat akurasi menjadi metrik yang menyesatkan karena model dapat memperoleh akurasi tinggi hanya dengan memihak kelas mayoritas [5], [6].

Kemajuan model berbasis *transformer* memberikan dasar yang kuat untuk mengatasi persoalan representasi. IndoBERT terbukti meningkatkan kualitas klasifikasi sentimen berbahasa Indonesia dibandingkan representasi berbasis frekuensi, baik pada ulasan umum [7], ulasan pelanggan berbasis aspek [8], maupun konten wisata [9]. Sejumlah penelitian lain membandingkannya secara langsung dengan algoritma konvensional pada data media sosial dan komentar YouTube [10], [11], [12], [13]. Namun keunggulan tersebut tidak selalu besar. Pada isu korupsi, IndoBERT hanya berada pada kisaran akurasi yang berdekatan dengan *Logistic Regression* dan *Linear SVM* [10]. Pada komentar YouTube bertema sensitif, akurasi *fine-tuning* IndoBERT dilaporkan berada di kisaran 74% [11]. Temuan bahwa penambahan data berlabel memberikan imbal hasil yang menurun setelah titik tertentu [14] juga menunjukkan bahwa perbaikan tidak dapat diharapkan datang dari volume data semata.

Dua arah perbaikan yang relevan telah diusulkan secara terpisah. Arah pertama menyangkut cara representasi kalimat dibentuk. Penggunaan *token* [CLS] sebagai satu-satunya ringkasan kalimat memperlakukan seluruh *token* secara implisit, sehingga mekanisme *attention* tambahan di atas lapisan akhir diusulkan untuk memberi bobot berbeda pada *token* yang lebih informatif [15], [16]. Arah kedua menyangkut *loss function*. *Focal loss* dan variannya dirancang untuk mengurangi kontribusi sampel mudah dan mengarahkan pelatihan pada sampel sulit, yang sebagian besar berasal dari kelas minoritas [17]. Pendekatan *contrastive* yang diawasi juga dilaporkan memperbaiki pemisahan representasi antarkelas pada klasifikasi teks pendek [18] dan pada data dengan distribusi kelas timpang [19]. Pada konteks berbahasa Indonesia, penanganan ketidakseimbangan lebih banyak diuji lewat penyeimbangan data seperti *random oversampling* [20] dan *oversampling* pada tugas berbasis aspek [21], bukan lewat modifikasi *loss function*.

Berdasarkan kajian terdahulu, persoalan yang belum terjawab bukan sekadar apakah IndoBERT mampu mengungguli algoritma berbasis TF-IDF, tetapi seberapa besar kontribusi tambahan setiap mekanisme ketika seluruh komponen dibandingkan pada kondisi eksperimen yang setara. *Attention pooling*, *supervised contrastive loss (SCL)*, dan *focal loss* telah dilaporkan pada konteks yang berbeda [15], [16], [17], [18], [19], tetapi evaluasi secara terpisah tidak dapat menunjukkan apakah suatu peningkatan berasal dari efek utama sebuah komponen, interaksi antarkomponen, atau variasi proses *fine-tuning*. Kekosongan tersebut menjadi lebih penting pada komentar YouTube berbahasa Indonesia yang sangat timpang dan memiliki ambiguitas sasaran sikap, karena perubahan *Macro-F1* yang kecil dapat terutama berasal dari kelas minoritas. Oleh karena itu, penelitian ini menggunakan rancangan faktorial penuh $2 \times 2 \times 2$ untuk mengisolasi kontribusi *attention pooling*, *SCL*, dan *dynamic adaptive focal loss* pada *fixed data split* dan beberapa *training seed*.

Penelitian ini menjawab tiga pertanyaan. Pertama, seberapa besar perbedaan performa antara *baseline TF-IDF*, *plain IndoBERT*, dan varian IndoBERT yang memperoleh komponen tambahan pada data yang sama. Kedua, apakah penambahan *attention pooling*, *SCL*, dan *dynamic adaptive focal loss* memberikan *marginal contribution* yang konsisten ketika komponen lain dikendalikan. Ketiga, pada kelas dan karakteristik komentar apa residual *disagreement* masih terjadi, serta sejauh mana reliabilitas *weak label* berbasis LLM membatasi interpretasi hasil. Kontribusi penelitian ini meliputi, pertama evaluasi *full-factorial* delapan varian IndoBERT pada *split* data yang tetap dengan tiga *training seed*. Kedua audit manusia terhadap 500 sampel label LLM menggunakan pedoman anotasi *target-aware*, dan yang terakhir analisis *paired marginal contribution*, performa per kelas, serta residual *disagreement* sehingga kontribusi komponen tidak disimpulkan hanya dari satu konfigurasi dengan skor tertinggi.

II. METODE

A. Alur Penelitian

Penelitian ini memakai pendekatan kuantitatif dengan metode eksperimen. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1, mulai dari pengumpulan data hingga analisis kesalahan.

B. Sumber Data dan Penyaringan Temporal

Pengumpulan awal melalui *YouTube Data API* menghasilkan metadata 75.453 komentar dari 54 video yang membahas kasus pemblokiran rekening nasabah Bank Mandiri. Tanggal 21 Agustus 2026 digunakan sebagai *case reference date* dan penyaringan dilakukan berdasarkan tanggal setiap komentar, bukan tanggal unggah video. Pemilihan atribut *comment_date* diperlukan karena sebuah video yang diunggah sebelum tanggal referensi tetap dapat menerima komentar setelah peristiwa.

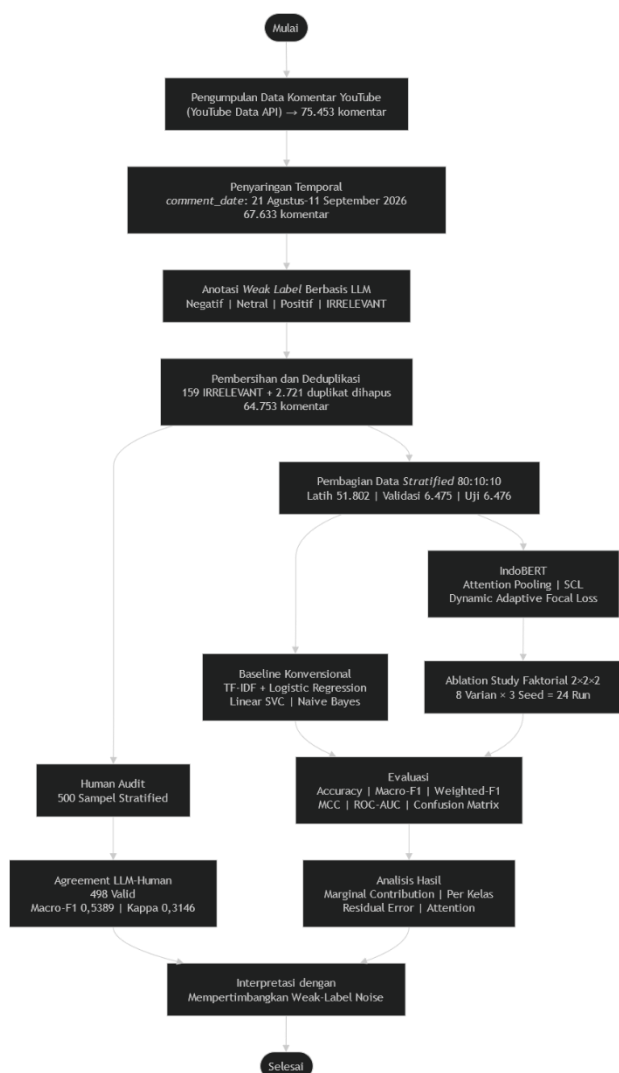
Setelah penyaringan berdasarkan *comment_date* pada rentang 21 Agustus 2026 hingga 11 September 2026, sebanyak 67.633 komentar dipertahankan. Selanjutnya, 159 komentar yang teridentifikasi sebagai *IRRELEVANT* dikeluarkan sehingga tersisa 67.474 komentar sebelum tahap pembersihan dan deduplikasi.

C. Praproses Teks dan Deduplikasi

Praproses dirancang seminimal mungkin karena model berbasis *transformer* membaca kapitalisasi, tanda baca, dan emoji sebagai sinyal. Langkah yang diterapkan meliputi normalisasi spasi berlebih, penghapusan baris dengan nilai kosong, dan penghapusan komentar yang menjadi kosong setelah pembersihan. Emoji, negasi, dan morfologi dipertahankan utuh. *Stemming* dan penghapusan *stopword* tidak diterapkan. Keputusan ini sejalan dengan pelaporan bahwa emoji membawa muatan sentimen pada komentar YouTube berbahasa Indonesia [22].

Deduplikasi teks persis kemudian dilakukan dan membuang 2.721 komentar. Langkah ini bertujuan mencegah

teks yang sama muncul pada data latih sekaligus pada data uji. Dataset akhir berjumlah 64.753 komentar.



Gambar 1. Alur Penelitian Klasifikasi Sentimen

D. Anotasi, Pedoman Kelas dan Distribusi Kelas

Seluruh korpus memperoleh weak label berbasis *Large Language Model (LLM)* menggunakan pedoman anotasi tertulis dengan Bank Mandiri sebagai target sikap. *NEGATIVE* didefinisikan sebagai kritik, ketidakpercayaan, kecaman, ajakan boikot atau pindah bank, serta evaluasi negatif lain yang diarahkan kepada Bank Mandiri. *POSITIVE* mencakup dukungan, pembelaan, apresiasi, kepercayaan, atau penilaian positif terhadap tindakan dan respons Bank Mandiri. *NEUTRAL* mencakup informasi tanpa evaluasi, pertanyaan murni, komentar yang masih terkait kasus tetapi evaluasinya diarahkan kepada aktor lain, serta kasus ambigu yang tidak dapat ditentukan tanpa menebak. *IRRELEVANT* tidak diperlakukan sebagai kelas sentimen dan dikeluarkan sebelum pemodelan.

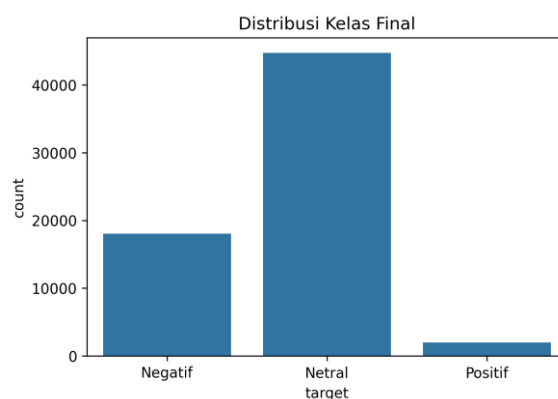
Sarkasme ditentukan berdasarkan makna yang dimaksud, bukan polaritas literal kata, karena ekspresi sarkastik dapat menunjukkan ketidaksesuaian antara makna literal dan sikap yang sebenarnya dimaksudkan [23]. Negasi dibaca sesuai cakupannya. Pada komentar dengan beberapa target, label ditentukan hanya dari sikap terhadap Bank Mandiri. Komentar yang memuji atau mengecam pihak lain tanpa mengevaluasi bank diberi label *NEUTRAL* dan dapat ditandai sebagai *off_target*. Emoji tanpa teks evaluatif tidak cukup untuk menentukan polaritas. Jika polaritas terhadap bank tetap tidak dapat ditentukan secara meyakinkan, anotator menggunakan *NEUTRAL* daripada menebak kelas positif atau negatif.

Untuk mengukur kualitas *weak label*, dilakukan human audit terhadap 500 komentar dengan *disproportionate stratified random sampling* berdasarkan label LLM menggunakan seed 2026. Sampel terdiri dari 150 *NEGATIVE*, 200 *NEUTRAL*, dan 150 *POSITIVE*. Label LLM disembunyikan dari file anotasi untuk mengurangi *anchoring bias*. Karena sampling tidak mengikuti prevalensi alami, evaluasi audit melaporkan metrik tanpa bobot dan *design-weighted metrics* menggunakan *inverse sampling weight* per strata LLM. Dua komentar dinilai *IRRELEVANT* oleh anotator manusia sehingga 498 komentar digunakan untuk audit tiga kelas. Audit ini merupakan pemeriksaan kualitas anotasi, bukan independent model test set.

Pendekatan ini telah diterapkan pada penelitian sentimen komentar *YouTube* berbahasa Indonesia sebelumnya [24]. Distribusi kelas hasil anotasi ditampilkan pada Tabel I dan Gambar 2.

TABEL I
DISTRIBUSI KELAS DATASET

Kelas	Jumlah	Proporsi
Negatif	18.031	27,85%
Netral	44.712	69,05%
Positif	2.010	3,10%
Total	64.753	100,00%



Gambar 2. Distribusi Kelas Dataset Final

Rasio kelas mayoritas terhadap kelas minoritas mencapai 22,24. Nilai tersebut cukup timpang untuk membuat akurasi

kehilangan daya bedanya, sehingga *Macro-F1* dipakai sebagai metrik utama [5].

E. Pembagian Data

Dataset dibagi secara berstrata pada label dengan rasio 80:10:10, menghasilkan 51.802 data latih terdiri dari 14.425 berlabel negatif, 35.769 berlabel netral, dan 1.608 berlabel positif. Pada data validasi terdapat 6.475 data yang terdiri dari 1.803 berlabel negatif, 4.471 berlabel netral, dan 201 berlabel positif. Terakhir 6.476 data pada data uji yang terdiri dari 1.803 berlabel negatif, 4.472 berlabel netral, dan 201 berlabel positif. Pembagian data menggunakan fixed split seed 42. Selanjutnya, seluruh varian IndoBERT dilatih menggunakan tiga *training seed*, yaitu 42, 123, dan 3407. Seluruh varian IndoBERT menggunakan data latih, validasi, dan uji yang sama, sedangkan *baseline* konvensional dievaluasi pada *fixed split* yang sama sehingga perbandingan antar model dilakukan pada data uji yang identik. Distribusi pembagian data ditampilkan pada Tabel II

TABEL II
PEMBAGIAN DATASET MENJADI TIGA DATA BAGIAN

Keterangan	Negatif	Netral	Positif	Jumlah Data	Negatif (%)	Netral (%)	Positif (%)
Data Latih	14.425	35.769	1.608	51.802	27.85	69.05	3.1
Data Validasi	1.803	4.471	201	6.475	27.85	69.05	3.1
Data Uji	1.803	4.472	201	6.476	27.85	69.05	3.1
Total	18.031	44.712	2.010	64.753	27.85	69.05	3.1

F. Baseline Konvensional

Tiga *baseline* dibangun di atas representasi *TF-IDF*, yaitu *Logistic Regression*, *Linear Support Vector Classifier (Linear SVC)*, dan *Multinomial Naive Bayes*. *Vectorizer* ditempatkan di dalam *Pipeline* sehingga kosakata dan bobot *IDF* hanya dihitung dari data latih dan tidak terjadi kebocoran informasi dari data uji.

Representasi *TF-IDF* menggunakan *word-level unigram (ngram_range=(1,1))*, *lowercase=True*, normalisasi L2, dan tanpa pembatasan *max_features*. *Logistic Regression* menggunakan *max_iter=1000* dan *random_state=42*. *Linear SVC* menggunakan *random_state=42* dengan *parameter* lain bawaan, sedangkan *Multinomial Naive Bayes* menggunakan *parameter* bawaan. Tidak diterapkan *class weighting* eksplisit pada ketiga *baseline*.

G. Arsitektur Model Usulan

Model usulan dibangun di atas *checkpoint indobenchmark/indobert-base-p1* [25] dengan tiga komponen tambahan.

1) *Attention Pooling*: Representasi kalimat tidak diambil dari *token [CLS]*, melainkan dari agregasi berbobot seluruh *token*. Diberikan *hidden state* h_i pada posisi ke- i , skor

kepentingan dihitung memakai Persamaan (1) dan bobot dinormalisasi memakai Persamaan (2).

$$e_i = v^T \tanh(W h_i + b) \quad (1)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^L \exp(e_j)} \quad (2)$$

Representasi kalimat kemudian dibentuk lewat Persamaan (3), dengan L menyatakan panjang urutan *token*.

$$s = \sum_{i=1}^L \alpha_i h_i \quad (3)$$

Token yang berada di luar *attention mask* diberi skor yang sangat rendah sebelum normalisasi sehingga tidak berkontribusi pada representasi.

2) *Supervised Contrastive Loss (SCL)*: Representasi s diproyeksikan lewat lapisan tersembunyi, dinormalisasi menjadi z , lalu dipakai untuk menghitung kerugian kontrastif memakai Persamaan (4). Notasi $P(i)$ menyatakan himpunan sampel sekelas dengan sampel ke- i dalam satu batch, $A(i)$ menyatakan seluruh sampel lain dalam batch, dan τ menyatakan parameter temperatur.

$$\mathcal{L}_{SCL} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (4)$$

3) *Dynamic Adaptive Focal Loss*: *Loss function* klasifikasi memakai *focal loss* dengan parameter fokus yang berubah mengikuti tingkat keyakinan model. Nilai γ pada setiap *epoch* dihitung memakai Persamaan (5), dengan $\bar{c}$ menyatakan rata-rata keyakinan prediksi pada *epoch* sebelumnya.

$$\gamma = \text{clip}(\gamma_{base} + 2(1 - \bar{c}), \gamma_{min}, \gamma_{max}) \quad (5)$$

Focal loss dihitung memakai Persamaan (6), dengan p_i menyatakan probabilitas yang diberikan model pada kelas sebenarnya.

$$\mathcal{L}_{focal} = -(1 - p_t)^\gamma \log(p_t) \quad (6)$$

Mekanisme ini membuat model memberi penekanan lebih besar pada sampel sulit ketika keyakinan rata-rata masih rendah pada awal pelatihan, lalu melunak seiring meningkatnya keyakinan. *Total loss* merupakan kombinasi berbobot keduanya seperti pada Persamaan (7).

$$\mathcal{L} = \lambda_{focal} \mathcal{L}_{focal} + \lambda_{SCL} \mathcal{L}_{SCL} \quad (7)$$

H. Rancangan Ablation Study

Kontribusi ketiga komponen diuji lewat *ablation study* faktorial penuh $2 \times 2 \times 2$ sehingga menghasilkan delapan varian, mulai dari IndoBERT tanpa komponen tambahan hingga kombinasi ketiganya. Varian yang tidak memakai *attention pooling* memakai *token [CLS]*, dan varian yang tidak memakai *focal loss* memakai *cross entropy* standar.

Seluruh varian dilatih dengan konfigurasi dan *seed* yang sama serta dievaluasi pada data uji yang sama.

I. Konfigurasi Pelatihan

Konfigurasi pelatihan ditampilkan pada Tabel III. Seluruh *neural run* menggunakan GPU NVIDIA GeForce GTX 1070 Ti dengan *mixed precision* saat CUDA tersedia. *Fixed split seed* adalah 42, sedangkan *training seed* yang digunakan adalah 42, 123, dan 3407, sehingga total terdapat 24 *neural run*

TABEL III
KONFIGURASI PELATIHAN

Parameter	Nilai
Pretrained model	indobenchmark/indobert-base-p1
Fixed split seed	42
Training seeds	42, 123, 3407
Jumlah neural run	24
Max sequence length	128
Batch size	16
Maximum epoch	5
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0,01
Warmup ratio	0,1
Gradient clipping	1,0
Dropout	0,1
Temperature SCL (τ)	0,07
$\lambda_{\text{classification}} / \lambda_{\text{SCL}}$	0,7 / 0,3
γ_{base}	2,0
$\gamma_{\text{min}} / \gamma_{\text{max}}$	1,0 / 5,0
Checkpoint selection	Validation Macro-F1
Early-stopping patience	5

Checkpoint terbaik pada setiap *run* dipilih hanya berdasarkan *Macro-F1* data validasi. Dengan maksimum lima *epoch* dan *patience* lima *epoch*, seluruh 24 *run* mencapai *epoch* kelima, *early stopping* tidak menghentikan *run* sebelum batas maksimum.

J. Analisis Multi-Seed dan Statistik

Untuk setiap varian *neural*, performa test dilaporkan sebagai *mean* $\pm$ *sample standard deviation* dari tiga *training seed*. *Interval* kepercayaan 95% dihitung menggunakan distribusi t dengan $n=3$ dan digunakan untuk menggambarkan ketidakpastian eksperimen, bukan sebagai bukti inferensial yang kuat. *Marginal contribution* dihitung secara *paired* pada *seed* yang sama dengan mengurangi *Macro-F1* dua varian yang hanya berbeda pada satu komponen. Test set tidak digunakan untuk memilih representative *checkpoint*.

K. Metrik Evaluasi

Metrik yang dilaporkan meliputi accuracy, balanced accuracy, precision, recall, *Macro-F1*, *Weighted-F1*, *Matthews Correlation Coefficient (MCC)*, dan ROC-AUC, disertai *confusion matrix* dan metrik per kelas. *Macro-F1* dipakai sebagai metrik utama karena memberi bobot sama

pada setiap kelas sehingga kegagalan kelas minoritas tidak tertutupi oleh kelas mayoritas [5], [6]. *MCC* dilaporkan sebagai metrik pendamping karena memperhitungkan seluruh sel *confusion matrix*.

III. HASIL DAN PEMBAHASAN

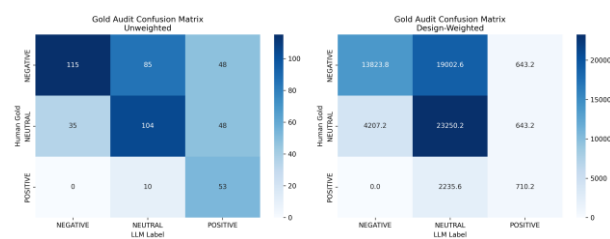
A. Audit Kualitas Anotasi LLM

Sebelum membandingkan model, *weak label LLM* diaudit terhadap anotasi manusia. Dari 500 sampel audit, dua komentar dinilai *IRRELEVANT* oleh satu anotator manusia, sehingga 498 komentar digunakan untuk evaluasi tiga kelas. Ringkasan agreement ditampilkan pada Tabel IV.

TABEL IV
AUDIT LABEL LLM TERHADAP HUMAN GOLD

Metrik	Unweighted	Design-weighted
Sampel audit	500	500
Human IRRELEVANT	2 (0,40%)	0,37%
Valid 3-class	498	498
Accuracy	0,5462	0,5857
Macro-Precision	0,5483	0,5483
Macro-Recall	0,6204	0,4938
Macro-F1	0,5389	0,4883
Cohen's kappa LLM-human	0,3146	0,2511

Audit menunjukkan *weak label LLM* masih mengandung *noise* yang substansial. Pada Gambar 3 *confusion matrix*, 85 komentar yang diberi label *NEGATIVE* oleh *human annotator* diprediksi sebagai *NEUTRAL* oleh LLM, sedangkan 48 komentar *human NEGATIVE* dan 48 komentar *human NEUTRAL* diberi label *POSITIVE* oleh LLM. Precision kelas *POSITIVE* terhadap *human gold* hanya 0,3557. *Metadata human* juga menandai 116 dari 500 komentar sebanyak 23,20% sebagai *off_target* dan 7 komentar sebanyak 1,40% sebagai sarkastik.



Gambar 3. Confusion Matrix Audit Label Llm Terhadap Human Gold

Temuan ini membatasi interpretasi seluruh hasil *downstream*. *Train*, *validation*, dan *test* utama tetap memakai *weak label LLM* sehingga skor model mengukur kemampuan mereproduksi struktur label tersebut, bukan estimasi langsung terhadap *human ground truth*. *Cohen's kappa* yang dilaporkan pada audit merupakan *agreement LLM-human*, bukan *inter-annotator agreement* antar manusia.

B. Performa Baseline Konvensional

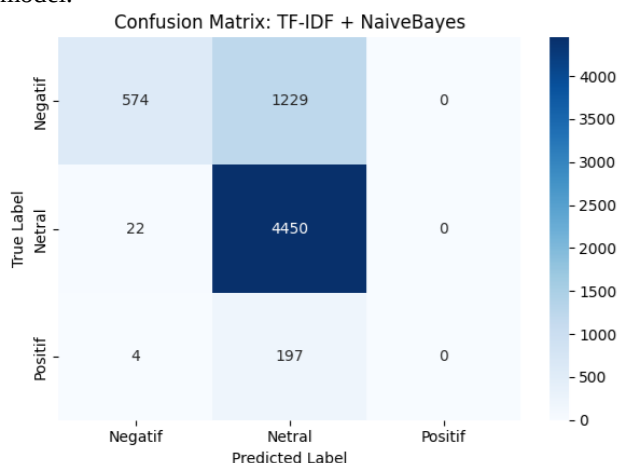
Hasil ketiga baseline pada *fixed test set* berlabel LLM ditampilkan pada Tabel V.

TABEL V
PERFORMA BASELINE KONVENSIONAL (TF-IDF) PADA DATA UJI (N = 6.476)

Model	Acc.	Macro-P	Macro-R	Macro-F1	Weight-F1	MC C
Logistic Regression	0,8848	0,8783	0,7017	0,7592	0,8789	0,7302
Linear SVC	0,8924	0,8558	0,7478	0,7905	0,8890	0,7500
Naive Bayes	0,7758	0,5713	0,4378	0,4459	0,7269	0,4514

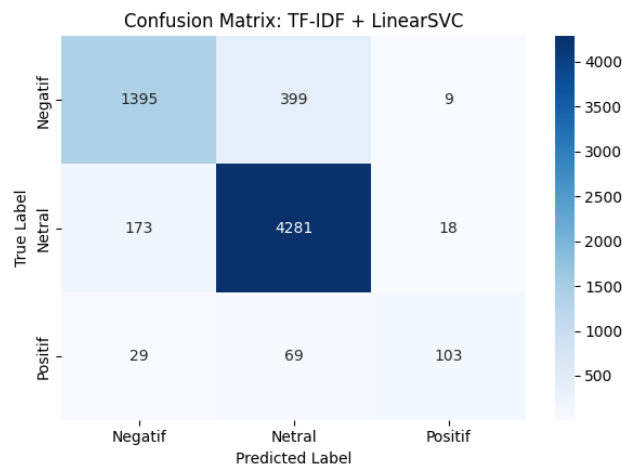
Jarak antara akurasi dan *Macro-F1* pada tabel tersebut memperlihatkan persoalan ketidakseimbangan secara langsung. *Logistic Regression* memperoleh akurasi 0,8848 tetapi *Macro-F1* hanya 0,7592, karena recall makro hanya 0,7017. Kasus paling ekstrem terjadi pada *Naive Bayes*. Model memperoleh akurasi 0,7758, tetapi performa tersebut didominasi oleh keberhasilan mengklasifikasikan kelas mayoritas Netral, sementara F1 kelas Positif bernilai 0,0000, yang menunjukkan bahwa tidak satu pun komentar positif berhasil dikenali. Temuan ini menegaskan bahwa akurasi tidak layak dipakai sebagai metrik tunggal pada data dengan rasio ketimpangan 22,24.

Untuk memperjelas temuan tersebut, rincian sebaran prediksi dari ketiga *baseline* konvensional dapat diamati secara langsung melalui *confusion matrix* masing-masing model.



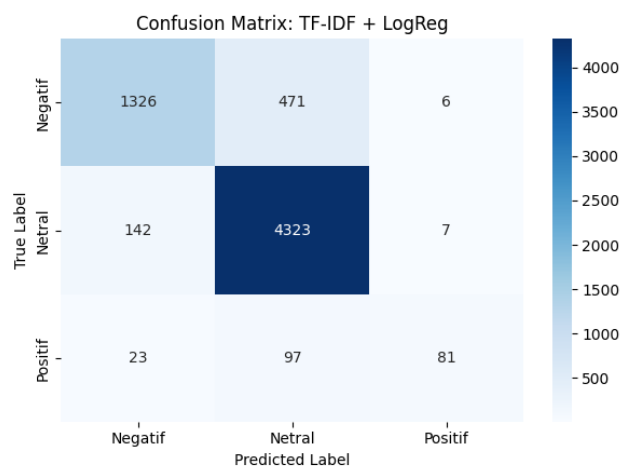
Gambar 4. Confusion Matrix Naive Bayes

Gambar 4 menunjukkan kegagalan *Naive Bayes* dalam mendeteksi kelas minoritas secara jelas, dari total 201 sampel aktual kelas Positif, tidak ada satu pun yang berhasil diprediksi dengan benar. Sebagian besar sampel kelas Positif sebanyak 197 sampel justru diprediksi ke dalam kelas mayoritas yaitu Netral, dan sisanya 4 sampel diprediksi sebagai Negatif.



Gambar 5. Confusion Matrix LinearSVC

Sementara itu, *confusion matrix* pada model *TF-IDF* dengan *Linear SVC* Gambar 5 menunjukkan performa pemisahan kelas yang paling baik di antara ketiga *baseline*. Model ini berhasil mengenali kelas minoritas dengan memprediksi 103 sampel kelas Positif secara tepat, dan juga mampu mengklasifikasikan 1.395 sampel kelas Negatif dengan benar. Meskipun menunjukkan peningkatan performa yang jelas dibandingkan *Naive Bayes*, bias terhadap kelas mayoritas tetap terlihat, yang ditandai dengan 69 sampel Positif dan 399 sampel Negatif yang keliru diprediksi sebagai kelas Netral.



Gambar 6. Confusion Matrix Logistic Regression

Pada *confusion matrix* model *TF-IDF* dengan *Logistic Regression* Gambar 6, performa yang dihasilkan berada di tengah-tengah antara *Naive Bayes* dan *Linear SVC*. Model ini hanya mampu memprediksi dengan benar 81 sampel dari kelas Positif, sementara porsi kesalahan yang cukup besar yaitu sebanyak 97 sampel Positif kembali salah diklasifikasikan sebagai Netral. Serupa dengan pola yang terjadi pada *baseline* lainnya, bias kelas mayoritas masih mendominasi pengambilan keputusan model, yang terlihat jelas dari 471 sampel aktual kelas Negatif yang gagal dipertahankan dan justru dilabeli sebagai kelas Netral.

Perbandingan F1 per kelas pada ketiga model *baseline* ditampilkan pada Tabel VI.

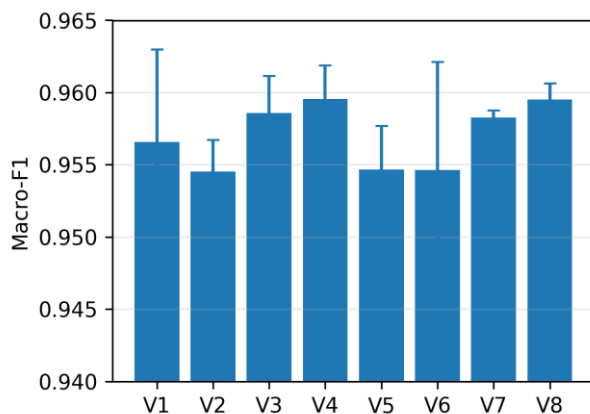
TABEL VI
PRECISION, RECALL, DAN F1-SCORE PER KELAS PADA BASELINE KONVENSIIONAL (TF-IDF)

Model	Kelas	Precision	Recall	F1
LogReg	Negatif	0,8893	0,7354	0,8051
LogReg	Netral	0,8839	0,9667	0,9234
LogReg	Positif	0,8617	0,4030	0,5492
LinearSVC	Negatif	0,8735	0,7737	0,8206
LinearSVC	Netral	0,9015	0,9573	0,9285
LinearSVC	Positif	0,7923	0,5124	0,6224
NaiveBayes	Negatif	0,9567	0,3184	0,4777
NaiveBayes	Netral	0,7573	0,9951	0,8601
NaiveBayes	Positif	0,0000	0,0000	0,0000

Pola yang konsisten pada ketiga *baseline* adalah kesenjangan yang lebar antara kelas Netral dan kelas Positif. *Linear SVC*, sebagai *baseline* terbaik, tetap kehilangan hampir setengah komentar positif dengan recall kelas tersebut hanya 0,5124. Representasi *TF-IDF* memperlakukan setiap kata sebagai fitur bebas tanpa konteks, sehingga komentar positif yang jumlahnya sedikit dan kosakatanya bertumpang tindih dengan komentar netral sulit dipisahkan.

C. Hasil Full-Factorial Ablation Multi-Seed

Hasil kedelapan varian IndoBERT ditampilkan pada Tabel VIII dalam urutan V1–V8. Perbandingan *mean Macro-F1* antarvarian ditampilkan pada Gambar 7.



Gambar 7. Mean ± Sd Macro-F1 Delapan Varian IndoBERT Pada Tiga Seed

Deskripsi delapan varian model disajikan pada Tabel VII untuk memperjelas konfigurasi setiap komponen dan menjaga Tabel VIII tetap ringkas.

TABEL VII
DESKRIPSI VARIAN MODEL

Varian	Attention	SCL	Dynamic Focal
V1	–	–	–
V2	✓	–	–
V3	–	✓	–
V4	–	–	✓
V5	✓	✓	–
V6	✓	–	✓
V7	–	✓	✓
V8	✓	✓	✓

TABEL VIII
HASIL UTAMA MULTI-SEED PADA WEAK-LABELLED TEST SET

V a r.	Accu racy	Macr o-P	Macr o-R	Macr o-F1	Weig hted-F1	MCC	Macr o AUC
V 1	0,982 6±0,0 020	0,960 9±0,0 063	0,952 4±0,0 066	0,956 6±0,0 064	0,982 6±0,0 020	0,960 8±0,0 044	0,992 4±0,0 003
V 2	0,982 6±0,0 005	0,957 1±0,0 035	0,952 1±0,0 075	0,954 5±0,0 022	0,982 6±0,0 005	0,960 8±0,0 012	0,994 3±0,0 016
V 3	0,983 6±0,0 011	0,959 6±0,0 024	0,957 6±0,0 036	0,958 6±0,0 026	0,983 6±0,0 011	0,963 2±0,0 024	0,994 3±0,0 011
V 4	0,984 0±0,0 002	0,962 3±0,0 049	0,956 8±0,0 001	0,959 5±0,0 023	0,984 0±0,0 002	0,963 9±0,0 004	0,993 9±0,0 007
V 5	0,982 1±0,0 007	0,958 5±0,0 012	0,951 0±0,0 055	0,954 7±0,0 030	0,982 0±0,0 007	0,959 6±0,0 016	0,992 9±0,0 004
V 6	0,982 8±0,0 025	0,954 0±0,0 126	0,955 6±0,0 020	0,954 7±0,0 075	0,982 8±0,0 024	0,961 2±0,0 055	0,992 7±0,0 037
V 7	0,983 2±0,0 004	0,958 2±0,0 012	0,958 4±0,0 005	0,958 3±0,0 005	0,983 2±0,0 004	0,962 2±0,0 008	0,994 9±0,0 012
V 8	0,983 7±0,0 005	0,960 4±0,0 042	0,958 7±0,0 036	0,959 5±0,0 011	0,983 7±0,0 005	0,963 4±0,0 012	0,993 9±0,0 009

Seluruh varian berbasis IndoBERT menghasilkan *Macro-F1* yang lebih tinggi dibandingkan *baseline* konvensional pada *weak-labelled test set*. *Mean Macro-F1* antarvarian IndoBERT berada pada rentang 0,9545–0,9595. V4, yaitu IndoBERT dengan *dynamic adaptive focal loss*, memperoleh *mean Macro-F1* tertinggi yang teramati sebesar $0,9595 \pm 0,0023$, sedangkan V8 yang menggabungkan ketiga komponen menghasilkan nilai hampir identik sebesar $0,9595 \pm 0,0011$. Dibandingkan *Linear SVC* sebagai *baseline* terbaik dengan *Macro-F1* 0,7905, peningkatan *mean Macro-F1* V4 mencapai sekitar 0,1690 secara absolut.

Nilai tertinggi berbeda menurut metrik evaluasi. V4 memperoleh *mean Accuracy* sebesar $0,9840 \pm 0,0002$, *Macro-Precision* sebesar $0,9623 \pm 0,0049$, *Weighted-F1* sebesar $0,9840 \pm 0,0002$, dan *MCC* sebesar $0,9639 \pm 0,0004$. V8 memperoleh *mean Macro-Recall* tertinggi sebesar 0,9587

$\pm 0,0036$, sedangkan V7 memperoleh *mean Macro ROC-AUC* tertinggi sebesar $0,9949 \pm 0,0012$. Oleh karena itu, perbedaan antarvarian tidak ditafsirkan hanya berdasarkan konfigurasi dengan *mean* tertinggi, tetapi dianalisis lebih lanjut melalui *paired marginal comparison* pada *training seed* yang sama.

D. Kontribusi Tiap Komponen

Marginal contribution dihitung secara *paired* pada *training seed* yang sama untuk seluruh 12 pasangan varian yang hanya berbeda pada satu komponen. Ringkasannya ditampilkan pada Tabel IX.

TABEL IX
PAIRED MARGINAL CONTRIBUTION MACRO-F1

Perbandingan	Komponen	Δ Macro-F1	SD	95% CI
V1→V2	Attention	-0,0020	0,0074	[-0,0204; 0,0163]
V1→V3	SCL	+0,0020	0,0053	[-0,0112; 0,0151]
V1→V4	Dynamic Focal	+0,0030	0,0070	[-0,0144; 0,0203]
V2→V5	SCL Att. ON	+0,0001	0,0017	[-0,0040; 0,0043]
V2→V6	Focal Att. ON	+0,0001	0,0074	[-0,0182; 0,0184]
V3→V5	Attention SCL ON	-0,0039	0,0055	[-0,0176; 0,0098]
V3→V7	Focal SCL ON	-0,0003	0,0030	[-0,0078; 0,0072]
V4→V6	Attention Focal ON	-0,0049	0,0090	[-0,0273; 0,0175]
V4→V7	SCL Focal ON	-0,0013	0,0025	[-0,0075; 0,0049]
V5→V8	Focal Att.+SCL ON	+0,0048	0,0040	[-0,0050; 0,0147]
V6→V8	SCL Att.+Focal ON	+0,0049	0,0078	[-0,0145; 0,0243]
V7→V8	Attention SCL+Focal	+0,0013	0,0015	[-0,0024; 0,0049]

Paired comparison memperlihatkan bahwa efek ketiga komponen bersifat kecil dan bergantung pada konfigurasi komponen lain. *Attention* bernilai negatif pada V1→V2 dan V3→V5, tetapi positif pada V7→V8. *SCL* bernilai positif pada V1→V3, V2→V5, dan V6→V8, tetapi negatif pada V4→V7. *Dynamic focal* juga menunjukkan *delta* positif pada beberapa kondisi dan sedikit negatif pada V3→V7. Seluruh 95% *confidence interval paired delta* melintasi nol sehingga tiga *seed* belum memberikan bukti bahwa satu komponen secara konsisten meningkatkan atau menurunkan *Macro-F1* pada seluruh kondisi.

Sebagai ringkasan faktorial deskriptif, rata-rata *main effect Macro-F1* adalah -0,00239 untuk *attention*, +0,00142 untuk *SCL*, dan +0,00189 untuk *dynamic focal*. *Dynamic focal* memiliki *main effect* rata-rata positif terbesar secara deskriptif, tetapi hasil tersebut tidak diperlakukan sebagai

klaim signifikansi statistik. Temuan ini juga menggantikan interpretasi *single-seed* sebelumnya yang menyimpulkan satu kombinasi tertentu sebagai sinergi terbesar. Penelitian yang melaporkan manfaat *SCL* umumnya bekerja pada kondisi data berlabel yang jauh lebih sedikit [18], [19].

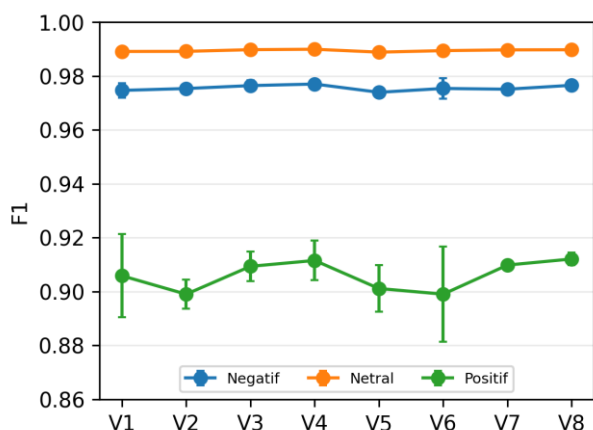
E. Performa Per Kelas

Nilai *Precision*, *Recall*, dan *F1-Score* per kelas untuk kedelapan varian ditampilkan pada Tabel X, sedangkan perbandingan *F1-Score* divisualisasikan pada Gambar 8.

TABEL X
MEAN $\pm$ SD PRECISION, RECALL, DAN F1-SCORE PER KELAS PADA DELAPAN VARIAN

Va r.	Kelas	Precision	Recall	F1-Score
V1	Negatif	0,9768 $\pm$ 0,0026	0,9726 $\pm$ 0,0027	0,9747 $\pm$ 0,0027
V1	Netral	0,9877 $\pm$ 0,0012	0,9906 $\pm$ 0,0010	0,9892 $\pm$ 0,0011
V1	Positif	0,9182 $\pm$ 0,0154	0,8939 $\pm$ 0,0160	0,9059 $\pm$ 0,0155
V2	Negatif	0,9793 $\pm$ 0,0030	0,9715 $\pm$ 0,0047	0,9754 $\pm$ 0,0009
V2	Netral	0,9874 $\pm$ 0,0024	0,9911 $\pm$ 0,0021	0,9892 $\pm$ 0,0002
V2	Positif	0,9046 $\pm$ 0,0100	0,8939 $\pm$ 0,0201	0,8990 $\pm$ 0,0054
V3	Negatif	0,9778 $\pm$ 0,0028	0,9752$\pm$0,0016	0,9765 $\pm$ 0,0020
V3	Netral	0,9892$\pm$0,0003	0,9905 $\pm$ 0,0009	0,9898 $\pm$ 0,0006
V3	Positif	0,9117 $\pm$ 0,0051	0,9071 $\pm$ 0,0104	0,9094 $\pm$ 0,0055
V4	Negatif	0,9810 $\pm$ 0,0005	0,9732 $\pm$ 0,0012	0,9771$\pm$0,0003
V4	Netral	0,9881 $\pm$ 0,0007	0,9919$\pm$0,0007	0,9900$\pm$0,0001
V4	Positif	0,9178 $\pm$ 0,0148	0,9055 $\pm$ 0,0000	0,9116 $\pm$ 0,0073
V5	Negatif	0,9784 $\pm$ 0,0039	0,9697 $\pm$ 0,0033	0,9740 $\pm$ 0,0013
V5	Netral	0,9867 $\pm$ 0,0008	0,9911 $\pm$ 0,0021	0,9889 $\pm$ 0,0008
V5	Positif	0,9104 $\pm$ 0,0070	0,8922 $\pm$ 0,0152	0,9011 $\pm$ 0,0087
V6	Negatif	0,9813$\pm$0,0027	0,9697 $\pm$ 0,0094	0,9754 $\pm$ 0,0038
V6	Netral	0,9875 $\pm$ 0,0024	0,9915 $\pm$ 0,0007	0,9895 $\pm$ 0,0012
V6	Positif	0,8933 $\pm$ 0,0377	0,9055 $\pm$ 0,0050	0,8990 $\pm$ 0,0177
V7	Negatif	0,9782 $\pm$ 0,0024	0,9721 $\pm$ 0,0031	0,9751 $\pm$ 0,0004
V7	Netral	0,9886 $\pm$ 0,0012	0,9909 $\pm$ 0,0007	0,9898 $\pm$ 0,0004
V7	Positif	0,9076 $\pm$ 0,0049	0,9121$\pm$0,0029	0,9098 $\pm$ 0,0010
V8	Negatif	0,9806 $\pm$ 0,0021	0,9726 $\pm$ 0,0006	0,9766 $\pm$ 0,0007

V8	Netral	0,9882±0,0004	0,9914±0,0011	0,9898±0,0004
V8	Positif	0,9123±0,0128	0,9121±0,0115	0,9121±0,0023



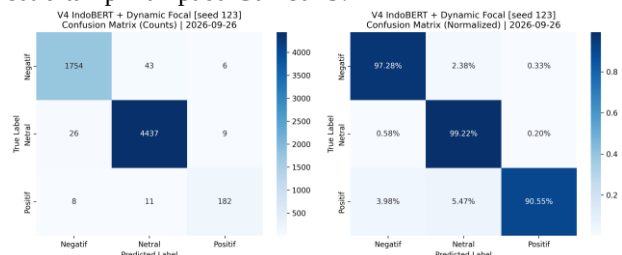
Gambar 8. Mean ± SD F1 Per Kelas Pada Tiga Training Seed

Perbedaan terbesar antarvarian tetap muncul pada kelas Positif. *Mean F1* Positif bergerak dari 0,8990 sampai 0,9121, sedangkan kelas Negatif berada sekitar 0,9740–0,9771 dan kelas Netral sekitar 0,9889–0,9900. V8 mencatat *mean F1* Positif tertinggi yang teramati sebesar $0,9121 \pm 0,0023$, sangat dekat dengan V4 sebesar $0,9116 \pm 0,0073$.

Plain IndoBERT V1 sudah mencapai *mean F1* Positif $0,9059 \pm 0,0155$. Dibandingkan *Linear SVC* dengan *F1* Positif 0,6224, lonjakan terbesar kembali berasal dari perpindahan representasi *TF-IDF* menuju representasi kontekstual IndoBERT. Komponen tambahan menghasilkan perubahan yang relatif lebih kecil dan tidak selalu searah pada setiap *seed*.

F. Confusion Matrix dan Analisis Kesalahan

Untuk analisis representatif, dipilih V4 *seed* 123 karena *seed* representatif ditentukan berdasarkan *validation Macro-F1*, bukan *test score*. *Confusion matrix* pada *weak-labeled test set* ditampilkan pada Gambar 9.



Gambar 9. Confusion Matrix Representatif V4 Seed 123

Representative *run* V4 *seed* 123 menghasilkan 103 disagreement dari 6.476 prediksi terhadap *weak label* atau sekitar 1,59%. Rincian per kelas ditampilkan pada Tabel XI.

TABEL XI
RINCIAN KESALAHAN VARIAN V4 SEED 123

Kelas	Kesalahan	Support	Error rate
Negatif	49	1.803	2,72%
Netral	35	4.472	0,78%
Positif	19	201	9,45%

Rate disagreement tertinggi tetap berada pada kelas Positif sebesar 9,45%. Arah terbesar adalah Negatif→Netral sebanyak 43 sampel, diikuti Netral→Negatif sebanyak 26 sampel. Pola tersebut konsisten dengan kesulitan membedakan evaluasi eksplisit terhadap Bank Mandiri dari komentar yang terutama menilai aktor lain.

Daftar *high-confidence disagreement* juga memperlihatkan kemungkinan *reference-label noise*. Sebagai contoh, komentar yang menyatakan ketakutan menabung di Mandiri atau ajakan memindahkan dana dapat memiliki *weak label* yang tidak sejalan dengan arah pragmatis teks, sementara komentar yang terutama membahas aparat dapat terklasifikasi berbeda karena target evaluasinya bukan bank. Karena itu, seluruh *sel off-diagonal* tidak otomatis diperlakukan sebagai kesalahan semantik model yang telah diverifikasi manusia.

Human gold audit memperkuat interpretasi tersebut: agreement LLM-human hanya mencapai *Macro-F1* 0,5389 dan kappa 0,3146. Dengan kata lain, residual analysis perlu membedakan model disagreement dari kemungkinan *weak-label disagreement*. Analisis kategori seperti *target ambiguity*, *off-target*, *sarcasm*, *implicit stance*, *mixed sentiment*, dan *likely weak-label error* merupakan arah adjudikasi lanjutan yang lebih tepat. Sarkasme secara khusus menjadi tantangan karena polaritas literal suatu ujaran tidak selalu merepresentasikan sikap yang sebenarnya dimaksudkan [23].

Hasil ini juga menjelaskan mengapa skor test terhadap *weak label* yang sangat tinggi tidak dapat disamakan dengan *human-grounded sentiment accuracy*. Keterbatasan anotasi menjadi bagian dari *error budget* penelitian, bukan hanya keterbatasan model.

Lima contoh *high-confidence residual disagreement* pada representative *run* V4 *seed* 123 ditampilkan pada Tabel XII. Contoh dipilih untuk mencakup beberapa kelas *weak label* dan arah *disagreement*, sehingga tidak hanya merepresentasikan satu pola kesalahan.

TABEL XII
CONTOH HIGH-CONFIDENCE RESIDUAL DISAGREEMENT PADA V4 SEED 123

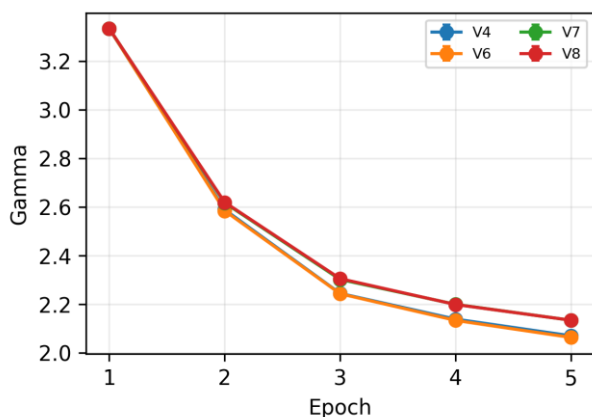
Komentar	True Label	Predicted Label	Confidence
Bank kocak dengan Polisi yg konyol adalah blunder terparah yg mlh bikin keadaan tambah runyam	Negatif	Netral	0,9984
Masyarakat pasti takut..nabung di mandiri..karena mereka bisa seenak jidat	Negatif	Netral	0,9977

memblokir hasil jirihpayah yang di tabung...pindahkan semua uang seluruh masyarakat..dari mandiri..			
Aparat Sudah Benar, menggalang dana dari Masyarakat Harus Ijin dan Tdk boleh pakai Rekening Peibadi	Positif	Netral	0,9953
Alhamdulillah.... Sy & keluarga serentak tutup rekening Mandiri 🙏🙏	Positif	Negatif	0,9846
mimpi apa mandiri..pas ketiban sial..tolol sih mau aj diperintah polisi ☐☐☐	Netral	Negatif	0,9842

G. Analisis Keyakinan Prediksi dan Dinamika Gamma

Dynamic adaptive focal loss memakai *true-class confidence epoch* sebelumnya untuk menentukan *focusing parameter*. Analisis *confidence* di bagian ini digunakan untuk memeriksa perilaku mekanisme adaptif, bukan sebagai analisis kalibrasi probabilitas. Penelitian ini tidak menghitung *expected calibration error*, *Brier score*, maupun *reliability diagram* sehingga tidak menyimpulkan bahwa model terkalibrasi.

Pada seluruh varian *focal* dan seluruh *seed*, γ dimulai sekitar 3,3333 karena epoch pertama menggunakan *prior confidence* 1/3. Seiring *confidence training* meningkat, nilai γ turun mendekati kisaran sekitar 2,06–2,14 pada epoch kelima. Pola multi-seed ditampilkan pada Gambar 10. Perubahan ini menunjukkan implementasi adaptif berjalan sesuai formulasi, tetapi bukan *sensitivity analysis* karena γ_{base} , γ_{min} , γ_{max} , dan koefisien adaptasi tidak dibandingkan pada beberapa konfigurasi.

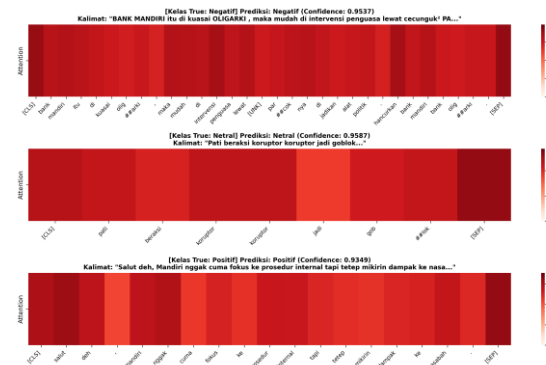


Gambar 10. Dinamika Mean $\pm$ SD Gamma Pada Varian Focal Dan Tiga Seed

H. Interpretasi Bobot Attention

Visualisasi bobot attention pada beberapa komentar uji ditampilkan pada Gambar 11. Hasil *direct comparison* terhadap *pooling* standar tidak menunjukkan peningkatan

attention yang konsisten. Pada $V1 \rightarrow V2$, *paired delta Macro-F1* adalah $-0,0020 \pm 0,0074$, sedangkan pada konteks *SCL* dan *focal* efeknya berubah arah. Karena itu, mekanisme *attention* tidak diklaim sebagai penyebab tunggal peningkatan performa. Interpretasi yang lebih tepat adalah bahwa kontribusinya bersifat *context-dependent* dan harus dibaca bersama interaksi komponen lain dalam *factorial design*.



Gambar 11. Visualisasi Bobot Attention Pada Komentar Uji

Meskipun *attention pooling* tidak menghasilkan peningkatan *Macro-F1* yang konsisten secara kuantitatif, visualisasi bobot *attention* tetap dapat digunakan untuk mengamati bagaimana model mendistribusikan perhatian terhadap token pada contoh komentar dari masing-masing kelas. Gambar 11 memperlihatkan distribusi bobot attention pada contoh komentar Negatif, Netral, dan Positif. Perbedaan intensitas antartoken menunjukkan bahwa mekanisme *attention pooling* memberikan kontribusi yang tidak seragam terhadap representasi kalimat. Namun, visualisasi ini bersifat deskriptif dan tidak dapat digunakan sebagai bukti kausal bahwa token dengan bobot lebih tinggi menjadi penyebab keputusan klasifikasi atau peningkatan performa model.

I. Keterbatasan

Penelitian ini memiliki beberapa keterbatasan. Pertama, seluruh *train*, *validation*, dan *test* utama menggunakan *weak label LLM*. *Human* audit 500 sampel menghasilkan *Macro-F1 LLM-human* 0,5389 dan *Cohen's kappa* 0,3146, sedangkan *design-weighted Macro-F1* 0,4883 dan *kappa* 0,2511. Dengan demikian, performa model sekitar 0,95 pada *test set* tidak dapat diperlakukan sebagai estimasi langsung terhadap *human ground truth*. Kedua, *human* audit dilakukan oleh satu anotator sehingga *inter-annotator agreement* antar-manusia belum tersedia. Ketiga, *gold sample* dibuat dari keseluruhan dataset untuk audit kualitas dan bukan dipisahkan sebagai *independent human-gold test set* sebelum *training*. Keempat, eksperimen *neural* menggunakan tiga *training seed* sehingga *statistical power* masih terbatas. Kelima, *baseline TF-IDF* tidak menggunakan *class weighting* dan tidak menjalani *hyperparameter search* khusus untuk menangani ketidakseimbangan kelas. Keenam, *exact duplicate* telah dihapus sebelum *splitting*, tetapi *near-duplicate* berbasis kemiripan tinggi belum dieliminasi secara eksplisit. Ketujuh,

dynamic adaptive focal loss hanya diuji pada $\gamma_{\text{base}}=2,0$, $\gamma_{\text{min}}=1,0$, dan $\gamma_{\text{max}}=5,0$ sehingga *sensitivity* terhadap *parameter* tersebut belum diketahui. Kedelapan, cakupan temporal dibatasi pada komentar yang dipublikasikan pada rentang 21 Agustus 2026 hingga 11 September 2026, sehingga hasil tidak merepresentasikan dinamika sentimen di luar periode pengumpulan tersebut.

IV. KESIMPULAN

Penelitian ini mengevaluasi *attention pooling*, *supervised contrastive loss*, dan *dynamic adaptive focal loss* pada IndoBERT melalui full-factorial ablation $2 \times 2 \times 2$ dengan *fixed train-validation-test split* dan tiga training *seed*. Pada *weak-labeled test set*, seluruh varian IndoBERT menghasilkan *Macro-F1* yang jauh lebih tinggi dibandingkan *baseline TF-IDF*. Plain IndoBERT memperoleh mean *Macro-F1* $0,9566 \pm 0,0064$, sedangkan *mean* tertinggi yang teramati diperoleh V4 dengan *dynamic adaptive focal loss* sebesar $0,9595 \pm 0,0023$. V8 yang menggabungkan ketiga komponen memperoleh nilai hampir identik sebesar $0,9595 \pm 0,0011$.

Paired marginal analysis menunjukkan bahwa efek *attention pooling*, *SCL*, dan *dynamic focal* bergantung pada konfigurasi komponen lain. Seluruh 95% *confidence interval* untuk *paired delta Macro-F1* melintasi nol pada tiga *seed* sehingga penelitian ini tidak mendukung klaim bahwa satu komponen secara konsisten meningkatkan atau menurunkan performa pada seluruh kondisi. Secara deskriptif, *dynamic focal* mempunyai rata-rata *main effect* positif terbesar, tetapi hasil tersebut belum cukup untuk dinyatakan signifikan. Kelas Positif tetap menjadi kelas tersulit, dengan *mean F1* sekitar 0,8990–0,9121 pada varian IndoBERT.

Human audit mengungkap keterbatasan yang lebih mendasar pada kualitas *weak label* seperti *agreement LLM-human* hanya menghasilkan *Macro-F1* 0,5389 dan *kappa* 0,3146 pada 498 sampel sentimen valid. Oleh sebab itu, hasil model harus difafsirkan sebagai performa terhadap *weak-labelled corpus*, bukan sebagai *human-grounded sentiment accuracy*. Penelitian lanjutan perlu memisahkan *human-gold test set* sebelum *training*, melibatkan lebih dari satu *annotator*, memeriksa *near-duplicate leakage*, serta melakukan *sensitivity analysis* terhadap *parameter dynamic focal* dan bobot *loss*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Bina Nusantara, BINUS Graduate Program (BGP), dan Research and Technology Transfer Office (RTTO) atas bimbingan, arahan, serta dukungan dalam penelitian ini. Penulis menggunakan Claude untuk anotasi dataset, ChatGPT dan Grammarly untuk membantu penyuntingan bahasa dan perbaikan tata bahasa naskah.

Ketersediaan Data

Data yang digunakan pada penelitian ini dapat diakses atau diunduh pada <https://doi.org/10.5281/zenodo.22769878>

DAFTAR PUSTAKA

- [1] R. Al Balawi, Y. Hu, and L. Qiu, "Brand crisis and customer relationship management on social media: Evidence from a natural experiment from the airline industry," *Information Systems Research*, vol. 34, no. 2, pp. 442–462, 2023.
- [2] K. Nandini and M. Rahardi, "Sentiment Analysis of Economic Policy Comments on YouTube Using Ensemble Machine Learning," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2607–2615, 2025.
- [3] I. T. Shandy and U. P. Sanjaya, "Comparative Analysis of Machine Learning and IndoBERT Models for Sentiment Analysis of YouTube Comments on the Free Nutritious Meals Program," *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 2391–2399, 2026.
- [4] A. P. Widyasari and M. Muljono, "Sentiment Analysis of YouTube Comments on the 2025 DPR RI Demonstration Using Machine Learning," *Journal of Applied Informatics and Computing*, vol. 10, no. 2, pp. 1729–1740, 2026.
- [5] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, 2024.
- [6] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices," *Remote Sens. (Basel)*, vol. 16, no. 3, p. 533, 2024.
- [7] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, 2022.
- [8] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024.
- [9] R. I. Perwira, V. A. Permedi, D. I. Purnamasari, and R. P. Agusdin, "Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, 2025.
- [10] 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.
- [11] M. Teddy Oswaria, T. Yusnitasaria, S. Wijaya, and J. M. Raya, "Sentiment Analysis of Indonesian YouTube Reviews About Lesbian, Gay, Bisexual, and Transgender (LGBT) using IndoBERT Fine Tuning," *LONTAR KOMPUTER*, vol. 15, no. 1, Apr. 2024.
- [12] A. S. Wijaya, A. Fahmi, and Y. T. C. Pramudi, "Sentiment Classification of Health Education YouTube Comments Using IndoBERT Embeddings with Logistic Regression and Naïve Bayes," *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 3005–3020, 2026.
- [13] S. Pradana and E. Kartikadarma, "A Comparison of Classical Machine Learning and IndoBERT on Sentiment Analysis of Danantara Program in X," *Journal of Applied Informatics and Computing*, vol. 10, no. 4, pp. 3954–3964, 2026.
- [14] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 22, 2025.
- [15] J. Seo, S. Lee, L. Liu, and W. Choi, "TA-SBERT: Token attention sentence-BERT for improving sentence representation," *IEEE Access*, vol. 10, pp. 39119–39128, 2022.
- [16] B. Chen, W. Peng, and J. Song, "MAI-CBert: Multidirectional Attention Interaction Construction-Bert Sentiment Sentence Representation," *IEEE Access*, vol. 10, pp. 70295–70304, 2022.
- [17] Z. Labd, S. Bahassine, and K. Housni, "AFL-BERT: Enhancing Minority Class Detection in Multi-Label Text Classification with Adaptive Focal Loss and BERT," *International Journal of*

- Advanced Computer Science & Applications*, vol. 16, no. 7, p. 496, 2025.
- [18] M. Wu, "Label knowledge-guided heterogeneous graph contrastive learning for semi-supervised short text sentiment classification," *J. Big Data*, vol. 12, no. 1, p. 227, 2025.
- [19] L. Qian, Q. Zuo, D. Li, and H. Zhu, "DGMSCl: A dynamic graph mixed supervised contrastive learning approach for class imbalanced multivariate time series classification," *Neural Networks*, vol. 185, p. 107131, 2025.
- [20] D. R. Alfinsyah and B. P. Hartato, "Evaluating the Impact of Random Over Sampling on IndoBERT Performance for Indonesian Sentiment Analysis," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3270–3282, 2025.
- [21] T. Aprilah and W. Herowati, "Enhancing aspect-based sentiment analysis via Hugging Face fine-tuned IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3821–3830, 2025.
- [22] M. N. F. Hidayat, D. D. Prasetya, and T. Widiyaningtyas, "Sentiment analysis of emoji and latinized Arabic in Indonesian YouTube comments: A LABERT-LSTM model," *Journal of Applied Engineering and Technological Science*, vol. 6, no. 2, pp. 1324–1341, 2025.
- [23] W. Chen, F. Lin, G. Li, and B. Liu, "A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities," *Neurocomputing*, vol. 578, p. 127428, 2024.
- [24] A. H. Nasution and A. Onan, "ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks," *Ieee Access*, vol. 12, pp. 71876–71900, 2024.
- [25] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857.