Evaluasi Kinerja IndoBERT dan N-Gram untuk Klasifikasi Umpan Balik Pengajaran

Yohanes[1], Hery Oktafiandi[2], Febriyanti Panjaitan[3], M.Fajar Ramadhan[4] , Winarnie,[5]

[1,2,3,4,5] *Program Studi Program Studi Teknik Informatika - Satu University, Indonesia*

Jl. Letda Abdul Rozak - Kec. Ilir Tim. II - Kota Palembang - Sumatera Selatan - Indonesia

yohanes@univ.satu.ac.id , hery.oktafiandy@univ.satu.ac.id , pebrianti.panjaitan@univ.satu.ac.id, m.ramadhan@univ.satu.ac.id, winarnie@univ.satu.ac.id

Abstrak—**Pendidikan adalah proses fundamental yang bertujuan untuk mencerdaskan kehidupan bangsa, dan harus dilaksanakan secara profesional oleh setiap guru. Salah satu pendekatan untuk meningkatkan kualitas pendidikan adalah dengan memperoleh umpan balik dari siswa, yang memungkinkan guru melakukan refleksi diri dan memperbaiki metode pengajarannya. Analisis sentimen membantu mengidentifikasi perasaan yang diungkapkan siswa dalam umpan balik mereka, apakah positif, netral, atau negatif. Selain itu, klasifikasi umpan balik digunakan untuk memperoleh wawasan mengenai berbagai aspek pengajaran. Penelitian ini mencapai tingkat akurasi terbaik sebesar 83,50% untuk klasifikasi umpan balik dan 88,35% untuk analisis sentimen menggunakan algoritma IndoBERT, yang sebelumnya melalui tahap preprocessing dengan metode N-Gram (N1 + N2), melampaui hasil penelitian sebelumnya. Berdasarkan model yang dikembangkan untuk menganalisis umpan balik siswa, dibuatlah sebuah aplikasi berbasis web yang dapat digunakan oleh kepala sekolah dan guru untuk menganalisis umpan balik secara lebih efisien dan efektif. Alat ini dapat menjadi sarana refleksi bersama untuk meningkatkan efektivitas pengajaran di dalam kelas.**

Kata Kunci— **Kualitas Pendidikan, IndoBERT, N-Gram, Analisis Sentimen, Umpan Balik Siswa.**

Abstract—**Education is a fundamental process aimed at educating the nation and must be carried out professionally by every teacher. One approach to improving the quality of education is by obtaining feedback from students, which enables teachers to reflect on and improve their teaching methods. Sentiment analysis helps identify the emotions expressed by students in their feedback, whether positive, neutral, or negative. In addition, feedback classification is used to gain insights into various aspects of teaching. This study achieved the best accuracy rates of 83.50% for feedback classification and 88.35% for sentiment analysis using the IndoBERT algorithm, preceded by preprocessing using the N-Gram (N1 + N2) method, outperforming previous studies. Based on the developed model for analyzing student feedback, a web-based application was created to assist principals and teachers in analyzing feedback more efficiently and effectively. This tool can serve as a medium for shared reflection to improve teaching effectiveness in the classroom.**

Keywords— **Education Quality, IndoBERT, N-Gram, Sentiment Analysis, Student Feedback.**

I. PENDAHULUAN

Pendidikan adalah proses yang bertujuan untuk mencerdaskan kehidupan bangsa, dan oleh karena itu, harus dilaksanakan secara profesional oleh setiap guru. Peran seorang guru dalam setiap proses pembelajaran sangatlah penting bagi masyarakat dan bangsa [1]. Kualitas pengajaran di sekolah memainkan peran yang sangat signifikan dalam menciptakan pengalaman belajar yang efektif dan bermakna bagi siswa, yang secara langsung berdampak pada prestasi dan pengalaman belajar mereka. Peningkatan kualitas guru berjalan seiring dengan peningkatan berkelanjutan terhadap kinerja siswa [2].

Menurut Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, guru yang hebat adalah mereka yang terus belajar dan menjadi teladan dalam pembelajaran sepanjang hayat [3]. Peningkatan kualitas pengajaran merupakan hal yang sangat penting dalam upaya peningkatan mutu pendidikan di Indonesia. Namun, evaluasi terhadap kualitas pengajaran sering kali bersifat subjektif dan sulit diukur secara objektif [4]. Terdapat kesenjangan dalam kualitas pengajaran antar guru, dan metode pengajaran yang efektif masih belum diterapkan secara merata.

Hasil Uji Kompetensi Guru (UKG) di Indonesia menunjukkan bahwa kurang dari 30% guru memperoleh nilai di atas 80, yang mengindikasikan tingkat kompetensi yang masih memprihatinkan [1]. Guru yang kompeten jumlahnya masih kalah banyak dibandingkan dengan guru yang kurang kompeten dalam mengajar. Tanpa evaluasi yang sistematis, upaya peningkatan kemampuan mengajar akan terhambat [5]. Penelitian ini bertujuan untuk mengatasi permasalahan tersebut dengan menggunakan algoritma analisis sentimen untuk memproses umpan balik siswa, sehingga dapat memberikan wawasan berharga bagi peningkatan kualitas pengajaran. Analisis sentimen menjadi alat yang berguna dalam memahami opini dan umpan balik [6].

Kualitas pengajaran seorang guru akan terlihat ketika ia melakukan refleksi diri dan berupaya menciptakan pengalaman belajar yang efektif dan berpusat pada siswa di dalam kelas. Kurangnya evaluasi dapat menghambat peningkatan kemampuan mengajar. Melalui penelitian ini, penulis berupaya memberikan kontribusi untuk mengatasi masalah tersebut dengan menggunakan algoritma analisis sentimen untuk

memproses umpan balik siswa dan mengubahnya menjadi wawasan yang dapat ditindaklanjuti guna mengevaluasi serta meningkatkan kualitas pengajaran [7].

Dalam penelitian ini, penulis menggunakan model deep learning dengan mengintegrasikan kecerdasan buatan ke dalam sistem evaluasi umpan balik untuk kinerja guru. Sistem ini terlebih dahulu melakukan klasifikasi terhadap umpan balik siswa, kemudian melakukan analisis sentimen menggunakan IndoBERT, dengan tujuan mempercepat pemahaman terhadap polaritas umpan balik.

BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model arsitektur yang dikembangkan dari transformer dengan menumpuk beberapa lapisan encoder untuk membentuk struktur baru [8]. BERT dirancang untuk memahami bahasa yang ambigu dalam kalimat dengan memanfaatkan kata-kata di sekitarnya untuk membangun makna [9]. Model BERT telah dilatih sebelumnya menggunakan dataset teks berbahasa Inggris dalam jumlah besar dari BooksCorpus dan Wikipedia [8]. Terdapat dua versi model BERT, yaitu BERT-base yang terdiri dari 12 lapisan encoder dan BERT-large yang memiliki 24 lapisan encoder [10].

Karena BERT awalnya dilatih menggunakan data berbahasa Inggris, maka dikembangkan model BERT yang disesuaikan dengan kondisi bahasa Indonesia, yaitu IndoBERT, untuk meningkatkan representasi bahasa dalam tugas-tugas NLP berbahasa Indonesia [8]. IndoBERT memperoleh F1 Score sebesar 84,13 dalam analisis sentimen, melampaui model lain seperti MBERT (76,58), MalayBERT (82,02), dan model pembelajaran mesin Naïve Bayes pada dataset yang sama skor F1 sebesar 70,95 [11].

IndoBERT merupakan model berbasis transformer dalam keluarga BERT yang dilatih sepenuhnya sebagai masked language model menggunakan kerangka kerja Hugging Face, mengikuti konfigurasi default BERT-base (uncased) [11]. Model ini terdiri dari 12 lapisan tersembunyi dengan dimensi masing-masing 768, memiliki 12 attention heads, dan lapisan tersembunyi feed-forward dengan dimensi 3.072. Kerangka kerja Hugging Face dimodifikasi untuk membaca aliran teks terpisah dari berbagai blok dokumen, dengan pelatihan menggunakan 512 token per batch. Kosakata IndoBERT mencakup 31.923 token dari lebih dari 220 juta kata yang dikumpulkan dari tiga sumber: Wikipedia (74 juta), sumber berita Indonesia (Kompas, Tempo, dan Liputan6) (55 juta), dan Indonesian Web Corpus (90 juta) [11].

Model N-gram adalah metode pre-processing yang digunakan untuk meneliti urutan n kata atau bunyi yang berurutan dalam teks atau ucapan [12]. Dalam analisis sentimen, model N-gram berperan penting dalam mengamati hubungan antar kata, yang pada gilirannya memperkuat deteksi sentimen. Model ini membantu mengklasifikasikan teks sebagai positif, negatif, atau netral berdasarkan kemunculan N-gram yang terkait dengan sentimen tertentu [13].

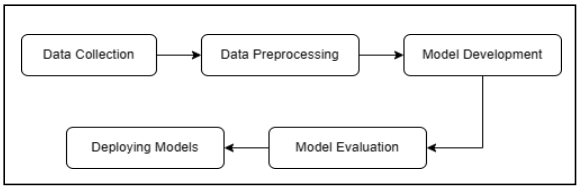
Penelitian sebelumnya oleh Wang et al. [14] berjudul "*Fuzzy-based Sentiment Analysis System for Analyzing Student Feedback and Satisfaction*" mengembangkan sistem evaluasi kepuasan terhadap kinerja guru menggunakan analisis sentimen dengan pendekatan Lexicon-Based Sentiment. Model ini mengatasi keterbatasan metode evaluasi siswa tradisional yang bergantung pada kuesioner konvensional dengan menyediakan wawasan tentang kinerja guru melalui proses otomatis. Sentimen dihitung menggunakan frekuensi kata dan polaritas untuk memperoleh skor sentimen keseluruhan dari umpan balik siswa. Penelitian ini menerapkan pendekatan berbasis *fuzzy logic* untuk menganalisis umpan balik dan kepuasan pelanggan. Data umpan balik siswa yang tersedia terlebih dahulu diproses menggunakan berbagai teknik *preprocessing*, seperti penghapusan *stopword*, *tokenization*, konversi huruf, dan koreksi ejaan. Langkah berikutnya adalah klasifikasi kata sentimen dan pergeseran polaritas, lalu dilakukan perhitungan skor sentimen keseluruhan. Sistem *fuzzy logic* kemudian digunakan untuk menganalisis umpan balik dan kepuasan pelanggan. Hasil penelitian menunjukkan performa yang sangat baik, di mana sistem yang diusulkan mengungguli metode klasifikasi mesin lainnya dalam hal akurasi, presisi, *recall*, dan *F-measure* [15].

II. METODE PENELITIAN

Penelitian ini bertujuan untuk mengembangkan studi sebelumnya oleh Wang et al., yang berjudul "*Fuzzy-based Sentiment Analysis System for Analyzing Student Feedback and Satisfaction*." Studi tersebut membahas sistem evaluasi kepuasan terhadap kinerja guru dengan pendekatan *fuzzy*-based *sentiment feedback* melalui klasifikasi kata opini dan polaritas dalam kalimat umpan balik dengan tingkat akurasi sebesar 82% [14].

Penelitian ini juga mengadopsi prinsip dari studi yang dilakukan oleh [16] berjudul "*An N-gram-Based BERT Model for Sentiment Classification Using Movie Reviews*," yang menggunakan dataset IMDB. Penerapan model N-gram berperan penting dalam menganalisis dan mempertimbangkan hubungan antar kata untuk memperkuat penentuan sentimen [13]. Pelatihan model dengan fitur N-gram memberikan indikasi yang baik tentang probabilitas kemunculan suatu kata setelah kata tertentu.

Dalam penelitian ini, penulis bertujuan untuk melanjutkan penelitian sebelumnya dengan menerapkan algoritma IndoBERT (*Indonesian Bidirectional Encoder Representations from Transformers*) untuk analisis sentimen terhadap umpan balik siswa. Pendekatan ini menggabungkan model IndoBERT dengan model N-gram untuk meningkatkan analisis dengan menangkap hubungan antar kata, sehingga meningkatkan akurasi deteksi sentimen. Penelitian ini juga bertujuan memberikan rekomendasi untuk peningkatan kualitas pengajaran berdasarkan wawasan yang diperoleh dari hasil analisis sentimen tersebut.



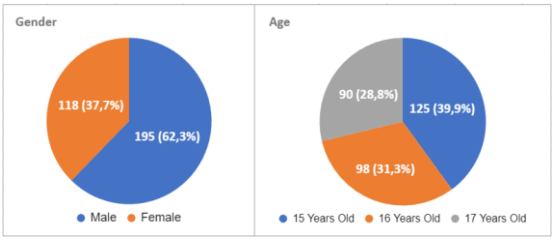
Gbr. 1 Tahapan Metode Penelitian

Pengumpulan data dilakukan untuk mencapai keseragaman data yang umum ditemukan pada studi kasus yang diangkat, sehingga model dapat belajar dan beradaptasi dengan dataset yang ada. Diharapkan model memiliki tingkat akurasi yang tinggi saat digunakan untuk melakukan prediksi dalam sistem umpan balik guru, baik dalam hal kategorisasi maupun analisis sentimen.

Pengumpulan data dilakukan dalam dua tahap dari berbagai responden siswa dengan beragam umpan balik terhadap beberapa guru untuk memastikan keberagaman dataset [17] Pengumpulan data dilakukan melalui pengisian Google Form berisi umpan balik mengenai metode pengajaran beberapa guru, yang diisi oleh siswa di SMK Methodist 2 Palembang. Setiap siswa diberikan 5 pertanyaan, namun tidak semua siswa menjawab seluruh pertanyaan. Beberapa pertanyaan yang disediakan dalam formulir umpan balik antara lain:

- Kesan Anda selama Miss mengajar
- Berikan kritik terhadap gaya mengajar Miss di kelas
- Berikan saran agar Miss dapat menjadi guru yang lebih baik di masa depan
- Bagaimana komunikasi Miss dalam menyampaikan informasi dan mengajar selama di sekolah?
- Metode pengajaran apa yang digunakan Miss yang sangat membantu dalam memahami materi?

Proses pelabelan dataset dilakukan menggunakan pendekatan *Crowdsourcing* yang melibatkan beberapa rekan guru dan kepala sekolah, berdasarkan penjelasan dari masing-masing label yang telah didefinisikan sebelumnya [18]. Pengumpulan data dilakukan dalam dua tahap dengan total 3 guru, melibatkan 313 responden dari siswa SMK Methodist 2 Palembang, dengan rincian yang ditunjukkan pada diagram lingkaran di Gambar 2 bawah ini:

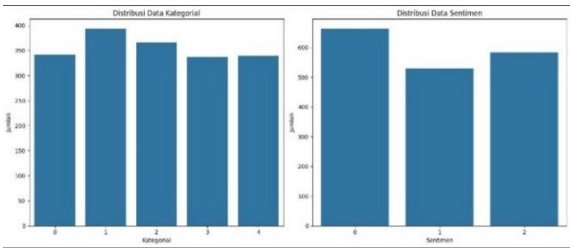


Gbr. 2 Persentase Responden

Dataset menunjukkan *imbalance label*, sehingga dilakukan augmentasi data secara manual dan menggunakan teknik *Generative AI* [19], yang merupakan tahap ketiga. Metode-metode ini disesuaikan agar selaras dengan konsep dari studi kasus yang diangkat serta pola-pola yang ditemukan dalam dataset yang telah dikumpulkan.

TABEL I
DATA TERKUMPUL

Tahap	Data	Kategori					Sentimen		
		0	1	2	3	4	0	1	2
Ke-1	506	150	159	17	103	77	185	161	160
Ke-2	1036	192	221	349	205	69	478	134	424
Ke-3	235	0	13	0	29	193	0	235	0
Jumlah	1777	342	393	366	337	339	663	530	584



(a) Kategorial (b) Sentimen

Gbr. 3 Perbandingan Label Dataset

Label Kategori:

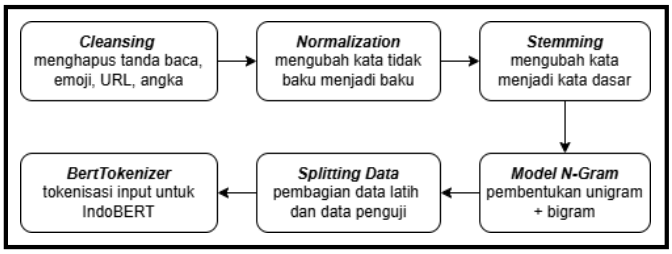
- [0] - Komunikasi / Penyampaian Materi
- [1] - Metode Pengajaran
- [2] - Pemberian Tugas / Ujian
- [3] - Pemahaman terhadap Siswa
- [4] - Tidak Terkategori

Label Sentimen:

- [0] - Positif
- [1] - Netral
- [2] – Negatif

Keterbatasan dataset dengan sumber daya rendah tidak menghambat penggunaan model IndoBERT dalam mengembangkan solusi untuk studi kasus saya yang bertujuan meningkatkan proses pembelajaran, dengan fokus pada peningkatan kinerja guru di dalam kelas. Penelitian yang dilakukan oleh [20] menunjukkan bahwa penggunaan metode transfer learning atau fine-tuning pada model deep learning memberikan kinerja yang cukup baik.

Dataset yang telah dikumpulkan dan diberi label melalui beberapa metode *preprocessing* untuk menghilangkan noise dari data, dengan harapan dapat meningkatkan dan mempertahankan kinerja akurasi model. Tahapan lengkapnya dapat dilihat pada Gambar 4. Pada langkah *preprocessing* ini, *stop words* tidak dihapus karena BERT dapat mempelajari stop words secara mandiri, sehingga keberadaannya tidak berpengaruh signifikan terhadap perhatian pada kata-kata yang ada [21].

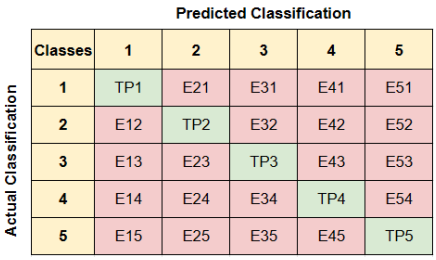


Gbr. 4 *Preprocessing Data*

Model IndoBERT yang digunakan dalam penelitian ini berasal dari Hugging Face, yaitu indobenchmark/indobert-base-p1. Selama proses pelatihan model, peneliti bertujuan untuk mengidentifikasi model dasar IndoBERT terbaik melalui *hyperparameter tuning* dan penentuan jumlah *epoch* yang

optimal [22], disesuaikan dengan dataset yang digunakan. Untuk proses *fine-tuning*, peneliti menggunakan Google Colab Pro bersama dengan pustaka PyTorch.

Evaluasi dilakukan menggunakan F1 Score, selaras dengan model IndoBERT lainnya pada huggingface.co/indobenchmark. Dengan dataset yang digunakan, tujuan evaluasi ini adalah untuk memperoleh kinerja model yang andal. Untuk menilai performa model yang dikembangkan, perhitungan dilakukan menggunakan confusion matrix guna mengukur akurasi, presisi, *recall*, dan *F1 Score* [23].



Gbr. 5 *Confusion Matrix*

III. HASIL DAN PEMBAHASAN

A. Hasil

Dalam penelitian ini, dilakukan fine-tuning pada IndoBERT dengan preprocessing N-Gram [13]. Eksperimen awal menggunakan hyperparameter yang mencakup *batch size* sebesar 32, Adam Optimizer, dan BertTokenizer, dengan *learning rate* 1e-4 dan 10 *epochs* [24]. Eksperimen dilakukan dengan berbagai implementasi N-Gram beserta kombinasinya sebelum dilakukan *hyperparameter tuning* [16]. Hasil dari eksperimen tersebut dapat dilihat pada Tabel 2.

TABEL III
HASIL PENGGUNAAN KOMBINASI N-GRAM

Ekstrasi Fitur	Klasifikasi Kategorial				Analisis Sentimen			
	Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
Tanpa N-Gram	77.99	78.67	77.99	77.96	83.50	86.79	83.50	84.56
Unigram	77.99	80.33	77.99	77.62	80.26	81.25	80.26	80.67
Bigrams	78.64	79.32	78.64	78.36	80.91	82.31	80.91	81.44
Trigrams	76.38	76.16	76.38	76.21	80.58	79.71	80.58	79.26
N1 + N2	**79.70**	**80.14**	**79.70**	**79.01**	**84.47**	**84.87**	**84.47**	**84.63**
N2 + N3	78.32	78.84	78.32	78.50	84.14	85.93	84.14	84.81
N3 + N1	78.67	79.57	78.67	78.92	81.23	86.84	81.23	82.84
Word2Vec	68.28	67.47	68.32	67.30	74.11	69.00	69.65	68.96

Berdasarkan Tabel 2, dapat diamati bahwa model ekstraksi fitur yang mencapai akurasi tertinggi selama proses *fine-tuning* model IndoBERT-base-p1 adalah kombinasi Unigram dan Bigram. Oleh karena itu, tahap pengujian selanjutnya dilakukan dengan menggunakan model ekstraksi fitur ini untuk menentukan akurasi terbaik selama *fine-tuning* [16]. Hasil akurasi terbaik diperoleh dengan kombinasi Unigram + Bigram, yang konsisten dengan temuan dalam penelitian [13].

Tahap berikutnya melibatkan pengujian model IndoBERT-base-p1, IndoBERT-lite-base-p1, BERT, dan LSTM menggunakan parameter konfigurasi yang sama yaitu 10 epoch, ukuran batch 32, dan learning rate 1e-4. Pengujian ini dilakukan bersamaan dengan penggunaan ekstraksi fitur N1 + N2, karena kombinasi tersebut menghasilkan akurasi terbaik pada eksperimen fine-tuning sebelumnya.

TABEL IIIII
PERBANDINGAN MODEL YANG TELAH DI FINE-TUNING

Model	Klasifikasi Kategorial				Analisis Sentimen			
	Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
Based Model (*Without Fine-Tuning*)	22.33	8.67	22.33	9.89	44.44	20.14	44.66	27.76
indobert-base-p1	**79.70**	**80.14**	**79.70**	**79.01**	**84.47**	**84.87**	**84.47**	**84.63**
indobert-lite-base-p1	74.76	76.01	74.76	73.96	82.20	83.15	82.20	82.55
BERT	79.29	79.31	79.25	79.25	82.52	84.48	82.52	83.20
LSTM	71.84	71.16	71.91	70.45	73.79	66.31	65.43	65.42

Berdasarkan hasil eksperimen yang ditampilkan pada Tabel 3, model IndoBERT-base-p1 menunjukkan kinerja terbaik di antara semua model yang diuji, baik pada tugas klasifikasi kategori maupun analisis sentimen. Model ini mencapai akurasi sebesar 79,70 dan F1 score sebesar 79,01 untuk klasifikasi kategori, serta akurasi sebesar 84,47 dan F1 score sebesar 84,63 untuk analisis sentimen.

Hal ini menunjukkan IndoBERT-base-p1 merupakan model yang paling efektif dalam menangani kedua tugas NLP tersebut setelah melalui proses pelatihan dengan pendekatan fine-tuning, dibandingkan dengan model lainnya sebagaimana terlihat pada Tabel 3. Eksperimen di atas juga menghasilkan akurasi analisis sentimen yang lebih baik dibandingkan penelitian [14], yang memperoleh akurasi sebesar 82% dengan pendekatan berbasis *fuzzy*, sedangkan eksperimen ini mencapai 84,47%. Meskipun perbandingan tersebut tidak sepenuhnya langsung karena adanya perbedaan dataset dan struktur data, hasil ini tetap dapat memberikan dasar perbandingan yang relevan.

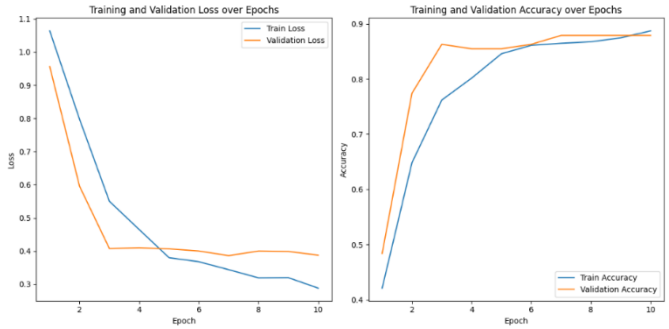
TABEL IVV
PERBANDINGAN MODEL YANG TELAH DI FINE-TUNING

#	Lr	*Batch Size*	Klasifikasi Kategorial				Analisis Sentimen			
			Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
1	1-e5	16	83.17	83.30	83.17	83.17	82.85	84.80	82.85	83.59
2	1-e5	32	**83.50**	**83.38**	**83.50**	**83.42**	85.11	86.05	85.11	85.51
3	3-e5	16	79.29	80.08	79.29	80.91	**88.35**	**87.85**	**88.35**	**87.80**
4	3e-5	32	80.65	83.19	80.65	80.91	87.90	88.11	87.90	87.64

Berdasarkan Tabel 4, percobaan ketiga dilakukan dengan menggunakan *learning rate* 3e-5, batch size 16, serta berbagai penyesuaian *hyperparameter tuning*.

Penyesuaian tersebut meliputi penambahan hidden dropout sebesar 0.5 dan penerapan scheduler menggunakan pustaka *get_linear_schedule_with*_warmup dari arsitektur transformer,

dengan nilai *num_warmup_steps* ditetapkan sebesar 0.1 [25]. Pada Gambar 6, percobaan ini berhasil mencegah proses pelatihan mengalami overfitting terhadap model, namun terjadi penurunan akurasi pada klasifikasi kategori sekitar ±4%. Penurunan akurasi ini kemungkinan disebabkan oleh tugas klasifikasi kategori yang melibatkan jumlah label lebih banyak dibandingkan klasifikasi analisis sentimen.



Gbr. 6 Analisis Grafik Akurasi Pelatihan dan Validasi Setelah Hyperparameter Tuning

B. Pembahasan

Hasil pengujian menunjukkan bahwa model IndoBERT-base-p1 dengan kombinasi Unigram dan Bigram (N1+N2) memberikan kinerja terbaik dengan akurasi 83,50% untuk klasifikasi kategorikal dan 88,35% untuk analisis sentimen. Temuan ini mengindikasikan bahwa pendekatan fine-tuning menggunakan model bahasa berbasis transformer mampu memahami konteks kalimat dalam bahasa Indonesia secara lebih mendalam dibandingkan pendekatan berbasis rule atau metode pembelajaran mesin tradisional.

Peningkatan performa model ini menunjukkan bahwa integrasi fitur N-Gram berperan penting dalam memperkuat hubungan antar kata yang sering muncul bersama. Hal ini sejalan dengan temuan [13] dan [16], yang menunjukkan bahwa N-Gram dapat meningkatkan sensitivitas model terhadap pola linguistik yang berulang dalam teks. Dengan demikian, N-Gram tidak hanya meningkatkan kemampuan klasifikasi semantik, tetapi juga membantu model mengurangi ambiguitas pada kalimat umpan balik siswa.

Jika dibandingkan dengan penelitian [14] yang menggunakan pendekatan *Fuzzy-Based Sentiment Analysis* dengan akurasi sebesar 82%, model yang dikembangkan dalam penelitian ini menunjukkan peningkatan sebesar 6,35% dalam akurasi analisis sentimen. GAP analysis ini menegaskan adanya peningkatan signifikan berkat penerapan transfer learning melalui IndoBERT yang telah dilatih pada korpus besar berbahasa Indonesia. Sementara penelitian Wang et al. masih terbatas pada pendekatan berbasis leksikon yang cenderung bersifat statis, penelitian ini memanfaatkan konteks dinamis antarkata yang diwakili oleh vektor embedding dari model transformer. Dengan demikian, hasil ini memperlihatkan adanya kebaruan dalam pendekatan analisis umpan balik pendidikan berbasis IndoBERT dan N-Gram.

Dari sisi manfaat praktis, sistem ini memungkinkan kepala sekolah dan guru untuk menganalisis persepsi siswa secara objektif dan cepat tanpa perlu membaca setiap umpan balik secara manual. Implementasi analisis berbasis AI ini dapat menjadi alat reflektif bagi guru untuk meningkatkan metode pengajaran sesuai aspek yang paling banyak dikomentari oleh siswa, seperti komunikasi, metode penyampaian materi, dan pemberian tugas. Temuan ini selaras dengan gagasan [26] mengenai pentingnya refleksi guru melalui tiga pendekatan *Reflection-for-Action*, *Reflection-in-Action*, dan *Reflection-on-Action* sehingga guru dapat memperbaiki kualitas pembelajarannya secara berkelanjutan.

IV. KESIMPULAN DAN SARAN

Model Fine-Tune IndoBERT dengan N1+N2 mencapai tingkat akurasi sebesar 83,50% dan F1 Score sebesar 83,42% untuk kategorisasi. Untuk analisis sentimen, model ini mencapai tingkat akurasi 88,35% dan F1 Score 87,80%. Tingkat akurasi ini menunjukkan potensi model dalam memberikan analisis yang efektif terhadap umpan balik siswa. Pada penelitian ini, fokus utama adalah pengukuran dan evaluasi kinerja model IndoBERT dengan pendekatan N-Gram hingga tahap analisis hasil. Implementasi model ke dalam platform berbasis web untuk digunakan oleh guru dan kepala sekolah akan menjadi arah penelitian selanjutnya. Melalui implementasi tersebut, diharapkan hasil klasifikasi dan analisis sentimen dari umpan balik siswa dapat memberikan wawasan yang lebih mendalam bagi guru untuk mengidentifikasi area perbaikan dan mengembangkan strategi pengajaran yang lebih efektif.

DAFTAR PUSTAKA

[1] A. H. Veirissa, “Kualitas Guru,” *Robbayana: Jurnal Pendidikan Agama Islam*, vol. 1, no. 2, hlm. 43–53, 2023, doi: 10.71029/robbayana.v1i2.28.

[2] R. Vagi, M. Pivovarova, dan W. Miedel Barnard, “Keeping Our Best? A Survival Analysis Examining a Measure of Preservice Teacher Quality and Teacher Attrition,” *J. Teach. Educ.*, vol. 70, no. 2, hlm. 115–127, Mar 2019, doi: 10.1177/0022487117725025.

[3] N. Makarim, “Nadiem Guru Hebat adalah Guru yang Terus Belajar.” Diakses: 14 Oktober 2025. [Daring]. Tersedia pada: https://mediaindonesia.com/humaniora/501261/nadiem-guru-hebat-adalah-guru-yang-terus-belajar

[4] R. J. Kreitzer dan J. Sweet-Cushman, “Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform,” *J. Acad. Ethics*, vol. 20, no. 1, hlm. 73–84, Mar 2022, doi: 10.1007/s10805-021-09400-w.

[5] R. Harrison *dkk.*, “Evaluating and enhancing quality in higher education teaching practice: a meta- review,” *Studies in Higher Education*, vol. 47, no. 1, hlm. 80–96, 2022, doi: 10.1080/03075079.2020.1730315.

[6] S. Meshram dan S. N. Damodar, “Feedback Sentiment Analysis for Teachers Recommendation using Data Mining,” *International Journal of Innovative Research*

in Science, vol. 8, No.10 November, 2019, doi: 10.15680/IJIRSET.2019.0810036.

[7] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, dan M. A. Wani, "Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study," *Applied Sciences (Switzerland)*, vol. 11, no. 9, 2021, doi: 10.3390/app11093986.

[8] J. Devlin, M. W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, hlm. 4171–4186, 2019.

[9] A. Vaswani *et al.*, "Attention Is All You Need," hlm. 1, Jun 2017, Diakses: 19 Oktober 2025. [Daring]. Tersedia pada: https://arxiv.org/pdf/1706.03762

[10] B. Juarto dan Yulianto, "Indonesian News Classification Using IndoBert," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 2, hlm. 454–460, 2023.

[11] F. Koto, A. Rahimi, J. H. Lau, dan T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, hlm. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[12] J. H. Computer, S. M. Honova, V. P. Computer, C. A. Setiawan, I. H. Parmonangan, dan Diana, "Sentiment Analysis of Skincare Product Reviews in Indonesian Language using IndoBERT and LSTM," *Proceeding - IEEE 9th Information Technology International Seminar*, 2023, doi: 10.1109/ITIS59651.2023.10420222.

[13] A. Tripathy, A. Agrawal, dan S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, March, hlm. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.

[14] Y. Wang, F. Subhan, S. Shamshirband, M. Z. Asghar, I. Ullah, dan A. Habib, "Fuzzy-based sentiment analysis system for analyzing student feedback and satisfaction," *Computers, Materials and Continua*, vol. 62, no. 2, hlm. 631–655, 2020, doi: 10.32604/cmc.2020.07920.

[15] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, dan W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, hlm. 47–58, Apr 2021, doi: 10.1016/J.FUTURE.2020.11.022.

[16] T. E. Trueman, A. K. Jayaraman, E. Cambria, G. Ananthakrishnan, dan S. Mitra, "An N-gram-Based BERT model for Sentiment Classification Using Movie Reviews," *International Conference on Artificial Intelligence and Data Engineering, AIDE 2022*, hlm. 41–46, 2022, doi: 10.1109/AIDE57180.2022.10060044.

[17] N. Cennamo *dkk.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors 2022, Vol. 22, Page 4157*, vol. 22, no. 11, hlm. 4157, Mei 2022, doi: 10.3390/S22114157.

[18] C. Shorten, T. M. Khoshgoftaar, dan B. Furht, "Text Data Augmentation for Deep Learning," *J. Big Data*, vol. 8, no. 1, hlm. 1–34, Des 2021, doi: 10.1186/S40537-021-00492-0/FIGURES/5.

[19] Y. Aliyu, A. Sarlan, K. Usman Danyaro, A. S. B. A. Rahman, dan M. Abdullahi, "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources," *IEEE Access*, vol. 12, hlm. 66883–66909, 2024, doi: 10.1109/ACCESS.2024.3398635.

[20] Y. Qiao, C. Xiong, Z. Liu, dan Z. Liu, "Understanding the Behaviors of BERT in Ranking," 2019, [Daring]. Tersedia pada: http://arxiv.org/abs/1904.07531

[21] N. Ettaik dan B. L. El Habib, "Hyperparameter Optimization in NLP Architectures," no. Bml 2021, hlm. 466–470, 2022, doi: 10.5220/0010736600003101.

[22] K. S. Nugroho, "Confusion Matrix untuk Evaluasi Model pada Supervised Learning | by Kuncahyo Setyo Nugroho | Medium." Diakses: 19 Oktober 2025. [Daring]. Tersedia pada: https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

[23] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, dan R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, hlm. 348–354, 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[24] P. Goyal *dkk.*, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour," Jun 2017, Diakses: 19 Oktober 2025. [Daring]. Tersedia pada: https://arxiv.org/pdf/1706.02677

[25] P. Kitchel, "3 Types of Reflection to Enhance and Improve Your Teaching." Diakses: 19 Oktober 2025. [Daring]. Tersedia pada: https://www.kdp.org/blogs/phil-kitchel/2022/05/06/3-types-of-reflection-to-enhance-and-improve-your