Sains Data Jurnal Studi Matematika dan Teknologi
ISSN 2986-903X

v4i1(2026) pp. 42-48
https://pub.nuris.ac.id/sainsdata

Dampak Pra-pemrosesan Teks pada Akurasi Analisis Sentimen Multi-Aspek Menggunakan IndoBERT

Muhammad Akmal Nabil Hibrizi[1], Doni Abdul Fatah.[2]
[1,2] Program Studi Sistem Informasi, Fakultas Teknik, Universitas Trunojoyo Madura, Bangkalan, Indonesia
Email: [1]220441100068@student.trunojoyo.ac.id, [2] doni.fatah@trunojoyo.ac.id

DOI : https://doi.org/10.52620/sainsdata.v4i1.301

ABSTRAK

Ulasan daring mengenai destinasi wisata pantai di Kabupaten Sumenep merupakan sumber data krusial, namun analisisnya terhambat oleh data yang tidak terstruktur, terutama kesalahan ketik (*typo*) yang signifikan menurunkan akurasi model. Penelitian ini berhasil mengatasi tantangan tersebut dengan membangun dan mengevaluasi beberapa skenario model analisis sentimen multi-aspek yang akurat menggunakan IndoBERT. Untuk memaksimalkan performa, penelitian ini menguji dampak dari dua inovasi utama yaitu sebuah modul koreksi ejaan cerdas yang mengkombinasikan Damerau-Levenshtein Distance dengan N-Gram, serta teknik teks augmentasi. Dengan kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM), penelitian menerapkan alur kerja sistematis mulai dari pra-pemrosesan hingga *fine-tuning* model. Hasil evaluasi perbandingan menunjukkan temuan yang menarik, model *baseline* (tanpa perlakuan pra-pemrosesan lanjutan) justru mencapai kinerja tertinggi dengan akurasi 96.12%. Sementara itu, model yang menggunakan koreksi ejaan dan augmentasi teks menunjukkan performa yang sedikit lebih rendah. Penelitian ini menghasilkan sebuah model yang sangat akurat dari data asli dan memberikan wawasan penting bahwa pada dataset tertentu, peforma model *Transformer* seperti IndoBERT sudah mampu menangani *noise* bahasa informal tanpa memerlukan pra-pemrosesan yang kompleks.

Kata Kunci: Analisis Sentimen Multi-Aspek, IndoBERT, Damerau-Levenshtein Distance, Pra-pemrosesan Teks, Augmentasi Teks.

PENDAHULUAN

Sektor pariwisata merupakan salah satu pilar fundamental dalam perekonomian Indonesia yang memiliki potensi besar untuk mendorong pertumbuhan regional. Kabupaten Sumenep, yang terletak di ujung timur Pulau Madura, dikenal dengan potensi wisata bahari yang melimpah, khususnya wisata pantai yang menjadi daya tarik utama bagi wisatawan lokal maupun mancanegara. Di era digital saat ini, di mana lebih dari 78% penduduk Indonesia terhubung dengan internet, proses pengambilan keputusan wisatawan sangat dipengaruhi oleh ulasan daring (*online reviews*) yang tersedia di berbagai platform digital seperti Google Maps. Ulasan-ulasan ini menjadi sumber data yang masif dan krusial (*User-Generated Content*) untuk memahami persepsi publik serta menjadi landasan evaluasi dan pengembangan strategis bagi para pengelola destinasi wisata (Af'idah et al., 2023).

Namun, pemanfaatan data ulasan ini dihadapkan pada tantangan yang signifikan. Data ulasan yang bersifat tidak terstruktur sering kali mengandung bahasa informal, singkatan, bahasa gaul, dan terutama kesalahan ketik (*typo*) yang unik untuk konteks Indonesia. Keberadaan noise kebahasaan ini secara drastis menurunkan akurasi model analisis sentimen otomatis (Rahman et al., 2024), karena model kecerdasan buatan gagal mengenali makna dan sentimen yang terkandung di dalamnya. Analisis manual untuk mengatasi masalah ini terbukti tidak efisien, memakan waktu, dan rentan terhadap bias.

Untuk mengatasi kompleksitas bahasa tersebut, bidang *Natural Language Processing* (NLP) telah mengembangkan model berbasis *Transformer* seperti *IndoBERT*, yang telah menunjukkan performa superior untuk berbagai tugas pemrosesan teks berbahasa Indonesia, termasuk analisis sentimen di domain pariwisata (Cahyaningtyas et al., 2021)(Jayadianti et al., 2022). Meskipun demikian, efektivitas model canggih seperti *IndoBERT* sangat bergantung pada dua faktor utama: kualitas dan kuantitas data pelatihan. Di sinilah letak celah penelitian (*research gap*) yang utama yaitu ketersediaan dataset ulasan wisata yang bersih, berlabel, dan bervolume besar untuk konteks lokal seperti wisata pantai di Sumenep masih sangat terbatas. Melatih model dengan data yang "kotor" dan terbatas berisiko tinggi menyebabkan *overfitting*, di mana model hanya mampu

v4i1(2026) pp. 42-48
https://pub.nuris.ac.id/sainsdata

menghafal data latih dan gagal melakukan generalisasi pada data baru, sehingga menghasilkan analisis yang tidak akurat dan tidak dapat diandalkan (Feng et al., 2021).

Untuk menjawab permasalahan ganda tersebut, penelitian ini bertujuan untuk meneliti dan mengevaluasi secara empiris dampak dari dua pendekatan pra-pemrosesan cerdas dengan mengintegrasikan model *IndoBERT* dengan dua pendekatan pra-pemrosesan cerdas. Pertama, untuk mengatasi masalah kualitas data, penelitian ini mengimplementasikan modul koreksi ejaan yang mengkombinasikan *Damerau-Levenshtein Distance* untuk mengidentifikasi kandidat kata yang benar dan model *N-Gram* untuk memvalidasi dan memilih kata yang paling sesuai secara kontekstual (Damerau, 1964). Kombinasi ini terbukti efektif untuk menangani kesalahan ketik yang umum dilakukan manusia, termasuk transposisi karakter. Kedua, untuk mengatasi keterbatasan kuantitas data, penelitian ini menerapkan teknik augmentasi teks seperti *Synonym Replacement* dan *Back Translation* untuk memperbanyak dan memperkaya variasi data pelatihan secara sintetis (Wei & Zou, 2019). Dengan mengintegrasikan kedua metode ini dalam satu alur kerja analisis sentimen multi-aspek, Dengan mengintegrasikan kedua metode ini dalam satu alur kerja analisis sentimen multi-aspek, penelitian ini akan menguji apakah pendekatan menyeluruh ini mampu membangun sebuah model yang lebih akurat dan tangguh dalam menghadapi data ulasan dunia nyata yang beragam dan tidak terstruktur.

TINJAUAN PUSTAKA

Analisis Sentimen Berbasis Aspek (ABSA)

Analisis sentimen level dokumen seringkali gagal menangkap nuansa dalam ulasan yang mengandung sentimen berlawanan terhadap aspek berbeda (misal: “pantainya indah, tapi toiletnya kotor”). ABSA bertujuan mengidentifikasi opini terhadap aspek spesifik ('keindahan', 'fasilitas'), sehingga memberikan wawasan yang lebih mendalam (Af’idah et al., 2023), dan dapat ditindaklanjuti (Azzahra & Wibowo, 2020).

Model Transformer dan IndoBERT

Arsitektur *Transformer* merevolusi NLP dengan mekanisme *self-attention* yang memungkinkannya memahami konteks secara bidireksional (Devlin et al., 2018). IndoBERT adalah varian BERT yang secara khusus di-*pre-trained* pada korpus masif Bahasa Indonesia, menunjukkan kinerja unggul untuk berbagai tugas NLP di Indonesia (Cahyawijaya et al., 2021)(Jayadianti et al., 2022)(Widansyah et al., 2024).

Koreksi Ejaan dan Teks Augmentasi

Untuk menangani *noise*, dua teknik diuji. Pertama, koreksi ejaan cerdas menggunakan Damerau-Levenshtein Distance yang efektif menangani transposisi karakter (Damerau, 1964), dikombinasikan dengan N-Gram untuk validasi kontekstual (Santoso et al., 2019) (Kokong et al., 2024). Kedua, Augmentasi Teks seperti *Synonym Replacement* dan *Back Translation* digunakan untuk memperbanyak data latih dan mengatasi ketidakseimbangan kelas (Wei & Zou, 2019) (Rahma & Suadaa, 2023).

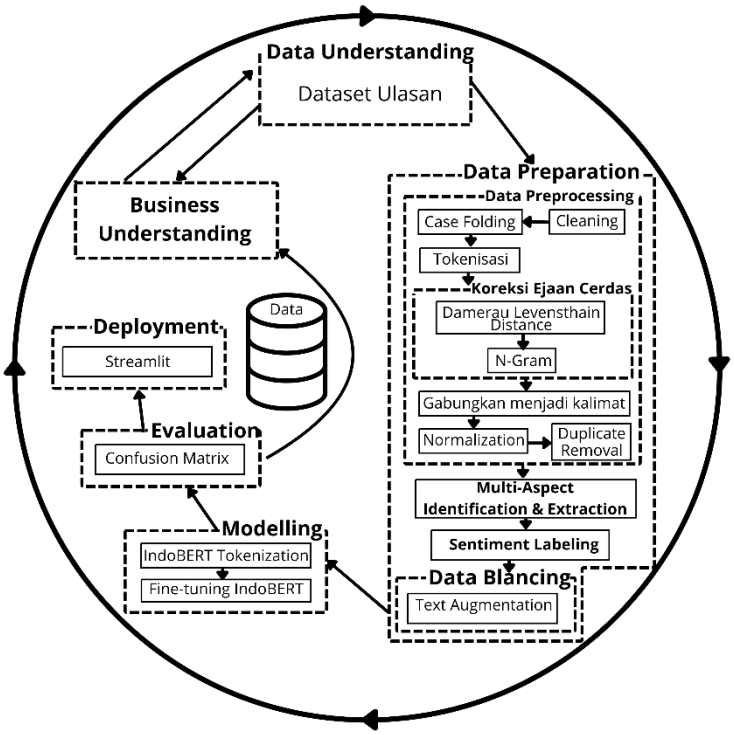
METODE PENELITIAN

Penelitian ini menggunakan pendekatan terukur dengan desain penelitian. Alur kerja penelitian mengadopsi kerangka *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang terstruktur.

Alur Penelitian

Penelitian ini mengadopsi kerangka *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang terstruktur (Wirth & Hipp, 2000). Alur kerja sistematis diilustrasikan pada Gambar 1.

Pada Gambar 1 mengilustrasikan alur kerja penelitian yang mengadopsi kerangka *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Proses ini dimulai dari pemahaman masalah (*Business Understanding*) dan pengumpulan data ulasan (*Data Understanding*). Data mentah kemudian melewati fase Persiapan Data yang mendalam, mencakup pembersihan, koreksi ejaan cerdas, ekstraksi multi-aspek, pelabelan, dan augmentasi teks. Dataset yang telah siap selanjutnya digunakan untuk melatih beberapa skenario model IndoBERT pada fase Pemodelan. Kinerja setiap model dievaluasi secara terukur (*Evaluation*) untuk menemukan pendekatan terbaik, yang kemudian direncanakan untuk diimplementasikan dalam sebuah aplikasi prototipe (*Deployment*). Kerangka kerja ini memastikan proses penelitian berjalan secara terstruktur dari awal hingga akhir.



Gambar 1 Diagram Alur CRISP-DM

Dataset

Data dikumpulkan melalui teknik *web scraping* dari ulasan publik di Google Maps untuk destinasi wisata pantai di Kabupaten Sumenep. Setelah melalui pra-pemrosesan (termasuk penghapusan duplikat), diperoleh 2.935 ulasan unik. Dari ulasan ini, diekstraksi 2.570 data poin berbasis aspek yang kemudian dilabeli secara semi-otomatis dan manual (Positif/Negatif). Aspek yang dianalisis adalah *ancillary* (layanan/harga), *attraction* (daya tarik), *amenities* (fasilitas), dan *accessibility* (aksesibilitas). Dataset kemudian dibagi menjadi 80% data latih dan 20% data uji (515 baris).

Desain Eksperimen

Empat skenario model dievaluasi untuk mengukur dampak setiap perlakuan:

1. Model A (Baseline)
   IndoBERT dilatih pada data asli yang hanya melalui pra-pemrosesan dasar. Tanpa menggunakan Damerau-Levenshtein Distance dengan N-Gram dan Teks Augmentasi.
2. Model B (Koreksi Ejaan)
   Data latih diperbaiki menggunakan modul koreksi ejaan cerdas berbasis Damerau-Levenshtein (Damerau, 1964) dan N-Gram (Santoso et al., 2019) sebelum pelatihan.
3. Model C (Lengkap)
   Data latih melalui kedua perlakuan yaitu Damerau-Levenshtein Distance dengan N-Gram dan Teks Augmentasi menggunakan teknik seperti *Back Translation* dan *Synonym Replacement* (Wei & Zou, 2019).

Evaluasi

Kinerja semua model dievaluasi pada data uji yang sama menggunakan metrik yaitu Akurasi, Presisi, *Recall*, dan F1-Score (Macro Average) (Powers, 2020).

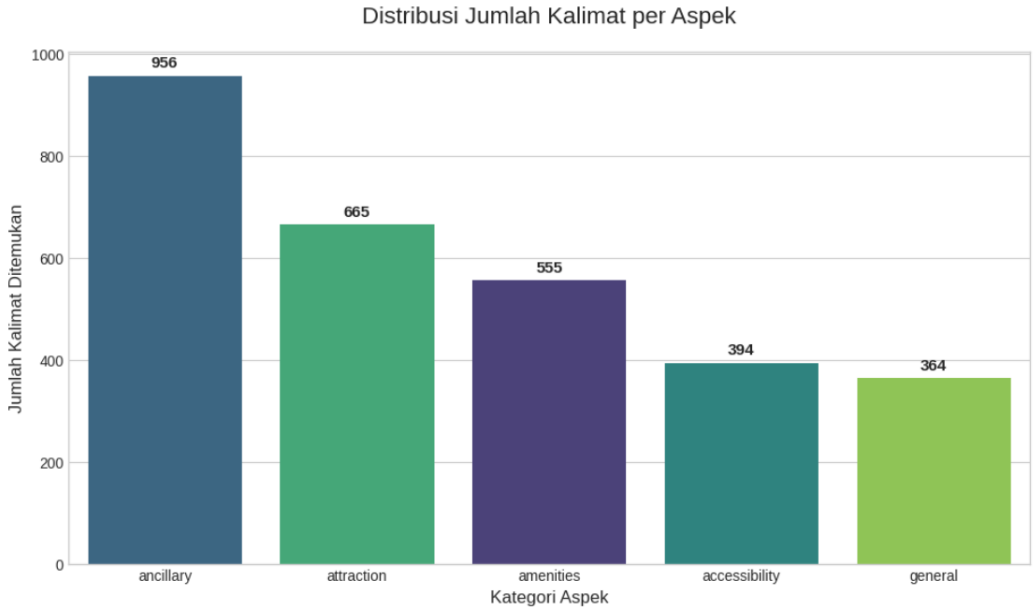
HASIL

Hasil Pra-pemrosesan dan Ekstraksi Aspek

Dari 3.211 ulasan mentah, proses pra-pemrosesan menghasilkan 2.935 ulasan bersih dilanjutkan ke tahap ekstraksi multi-aspek. Pada tahap ini, setiap ulasan dianalisis untuk mengidentifikasi kalimat atau frasa yang merujuk pada empat kategori aspek utama yaitu *ancillary* (layanan pendukung), *attraction* (daya

v4i1(2026) pp. 42-48
https://pub.nuris.ac.id/sainsdata

tarik), *amenities* (fasilitas), dan *accessibility* (aksesibilitas). Proses ekstraksi ini berhasil menghasilkan total 2.570 data poin yang relevan dengan keempat aspek tersebut, setelah 364 data poin yang dikategorikan sebagai general (tidak spesifik) disaring dan dihapus. Seperti yang ditunjukkan pada Gambar 2, aspek yang paling sering dibicarakan oleh pengunjung adalah aspek ancillary dengan hasil 956 kemunculan. Distribusi ini memberikan gambaran awal mengenai fokus utama dari ulasan pengunjung wisata pantai di Kabupaten Sumenep.

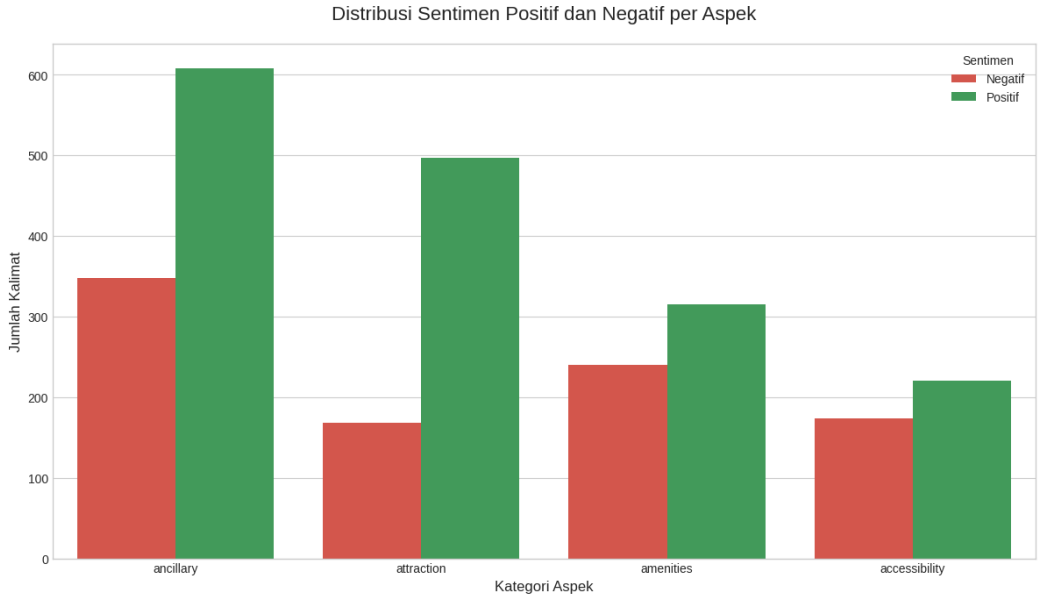


Gambar 2 Distribusi Jumlah Kalimat per Aspek

Hasil Pelabelan Sentimen dan Teks Augmentasi

Sebanyak 2.570 data poin berbasis aspek yang telah diekstraksi kemudian dilanjutkan ke tahap pelabelan sentimen. Proses ini dilakukan secara semi-otomatis, yang diawali dengan pelabelan otomatis menggunakan model *zero-shot classification* (joeddav/xlm-roberta-large-xnli), diikuti dengan verifikasi dan koreksi manual untuk memastikan akurasi dan kualitas setiap label (Positif/Negatif).

Hasil pelabelan pada keseluruhan dataset menunjukkan adanya ketidakseimbangan kelas sentimen yang signifikan pada beberapa aspek, seperti yang divisualisasikan pada Gambar 3. Terlihat bahwa ulasan untuk aspek *attraction* cenderung sangat positif, sementara ulasan untuk aspek lain memiliki distribusi yang lebih bervariasi.



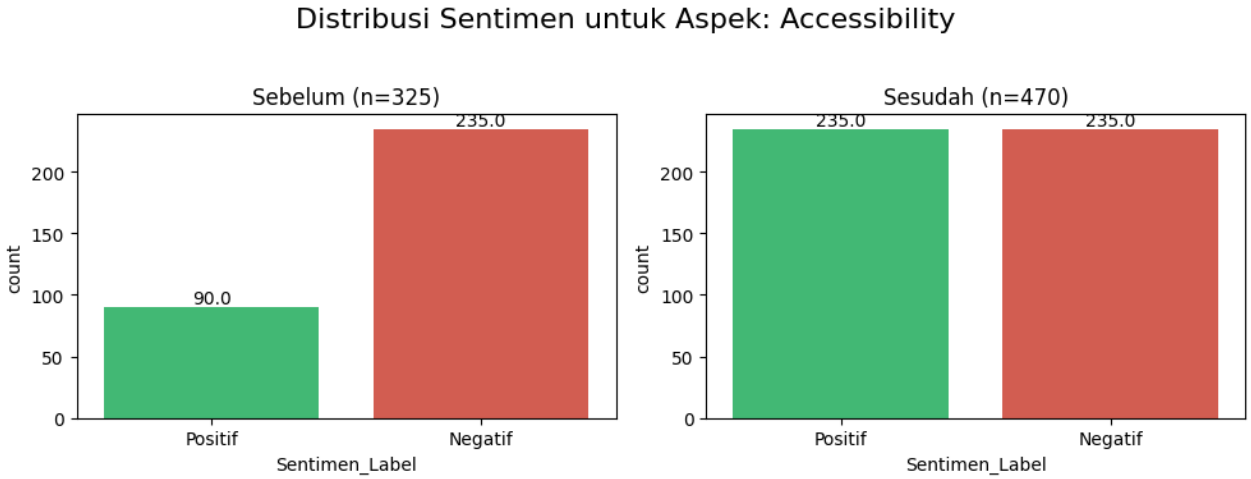
Gambar 3 Distribusi Sentimen Positif dan Negatif per Aspek

Untuk mempersiapkan pemodelan, data kemudian dibagi menjadi data latih (2.056 baris) dan data uji (514 baris). Ketidakseimbangan ini menjadi lebih jelas di dalam data latih, yang berisiko membuat model menjadi

v4i1(2026) pp. 42-48
https://pub.nuris.ac.id/sainsdata

bias. Sebagai contoh, pada data latih untuk aspek *Accessibility*, hanya terdapat 90 sampel bersentimen Positif dibandingkan dengan 235 sampel Negatif.

Untuk mengatasi masalah ini, teknik augmentasi teks meliputi *Synonym Replacement* dan *Back Translation* diterapkan secara strategis hanya pada kelas minoritas di dalam data latih. Tujuannya adalah untuk menyeimbangkan distribusi sentimen pada setiap aspek tanpa mengubah data uji. Proses ini berhasil menyeimbangkan data secara efektif. Gambar 4 menunjukkan perbandingan distribusi sentimen untuk aspek *Accessibility* sebelum dan sesudah augmentasi, di mana jumlah data untuk kelas Positif berhasil ditingkatkan dari 90 menjadi 235, setara dengan kelas Negatif.



Gambar 4. Distribusi Sentimen Aspek Accessibility Sebelum dan Sesudah Augmentasi

Melalui proses augmentasi yang ditargetkan pada setiap aspek, total data latih meningkat dari 2.056 menjadi 2.426 baris, sementara 514 baris data uji tetap dipertahankan dalam kondisi aslinya untuk memastikan proses evaluasi yang objektif dan andal.

Hasil Evaluasi Model

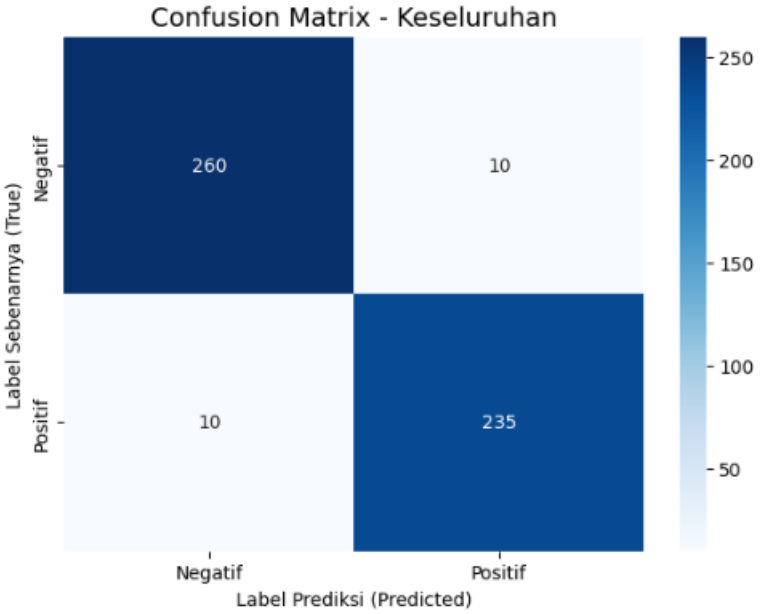
Evaluasi perbandingan dari keempat model pada 515 data uji menghasilkan temuan yang disajikan pada Tabel 1.

Tabel 1 Hasil Evaluasi Komparatif Model (Keseluruhan)

Model	Skenario	Akurasi	Presisi (Macro)	Recall (Macro)	F1-Score (Macro)
A	Model Baseline (tanpa Damerau-Levenshtein Distance dengan N-Gram dan teknik teks augmentasi)	0.9612	0.96	0.96	0.96
B	Modul Koreksi Ejaan Cerdas	0.9514	0.95	0.95	0.95
C	Model Koreksi Ejaan Cerdas dan Teks Augmentasi	0.9514	0.95	0.95	0.95

Secara tak terduga, Model A (Baseline) menunjukkan kinerja tertinggi di semua metrik. Analisis lebih dalam pada *confusion matrix* Model A (Gambar 5) menunjukkan kemampuannya yang seimbang dalam mengklasifikasikan kedua kelas sentimen.

Performa Model A juga sangat kuat ketika dianalisis per aspek, dengan F1-Score tertinggi pada aspek *Amenities* (0.98) dan terendah pada aspek *Accessibility* (0.87), namun tetap menunjukkan tingkat akurasi yang tinggi secara keseluruhan.



Gambar 5. Confusion Matrix - Model A (Baseline)

PEMBAHASAN

Temuan utama penelitian ini bahwa model *baseline* mengungguli model-model dengan pra-pemrosesan lanjutan berlawanan dengan hipotesis awal. Hasil ini mengindikasikan beberapa analisis penting.

Pertama, peforma bawaan dari IndoBERT kemungkinan besar menjadi faktor utama. Sebagai model *Transformer* (Devlin et al., 2018) yang telah di-*pre-trained* pada korpus data internet Indonesia yang sangat besar dan beragam (termasuk bahasa non-standar), IndoBERT telah mengembangkan kemampuan bawaan untuk memahami konteks kalimat bahkan ketika terdapat *noise* seperti *typo* atau singkatan(Pratama et al., 2025). Hal ini menunjukkan bahwa untuk model yang sudah sangat tangguh, kombinasi pra-pemrosesan yang mendalam mungkin tidak diperlukan.

Kedua, pra-pemrosesan lanjutan justru berpotensi memberikan dampak negatif kecil. Modul koreksi ejaan, meskipun akurat, dapat mengalami *over-correction*, yaitu mengubah istilah lokal atau slang yang memiliki makna kontekstual penting (Nur, 2021). Demikian pula, kalimat hasil augmentasi, meskipun menyeimbangkan dataset, mungkin terdengar kurang alami dan tidak menambah informasi makna baru yang signifikan bagi model yang sudah canggih (Feng et al., 2021), sehingga tidak memberikan manfaat tambahan.

Terakhir, hasil ini menggarisbawahi pentingnya validasi. Asumsi teori bahwa "lebih banyak pra-pemrosesan pasti lebih baik" terbukti tidak berlaku untuk kasus ini. Pendekatan yang lebih sederhana dengan mengandalkan kekuatan model *pre-trained* secara langsung ternyata menjadi strategi yang paling efektif.

KESIMPULAN

Penelitian ini menyimpulkan bahwa model IndoBERT menunjukkan ketangguhan yang sangat tinggi dalam menangani teks ulasan pariwisata berbahasa Indonesia yang informal, di mana model *baseline* (tanpa pra-pemrosesan Damerau-Levenshtein Distance dengan N-Gram) berhasil mencapai akurasi tertinggi sebesar 96.12%. Penerapan modul koreksi ejaan dan augmentasi teks tidak memberikan peningkatan performa pada kasus ini, mengindikasikan bahwa untuk model *Transformer* yang sudah tangguh, pra-pemrosesan yang kompleks tidak selalu menjamin hasil yang lebih baik. Pendekatan yang paling unggul adalah dengan langsung melakukan *fine-tuning* pada model IndoBERT setelah pra-pemrosesan dasar, membuktikan bahwa kekuatan arsitektur *Transformer* sudah memadai untuk memahami nuansa dan *noise* dalam data teks dunia nyata.

UCAPAN TERIMA KASIH

Penelitian ini terlaksana berkat bimbingan dari Bapak Doni Abdul Fatah, S.Kom., M.Kom., serta dukungan dari LPPM Universitas Trunojoyo Madura melalui program MBKM Penelitian.

DAFTAR PUSTAKA

Af'idah, D. I., Anggraeni, P. D., Rizki, M., Setiawan, A. B., & Handayani, S. F. (2023). Aspect-Based Sentiment Analysis for Indonesian Tourist Attraction Reviews Using Bidirectional Long Short-Term Memory. *JUITA : Jurnal Informatika*, *11*(1), 27. https://doi.org/10.30595/juita.v11i1.15341

Azzahra, S. A., & Wibowo, A. (2020). Analisis Ulasan Wisatawan. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(4), 737. https://doi.org/10.25126/jtiik.202071907

Cahyaningtyas, S., Hatta Fudholi, D., & Fathan Hidayatullah, A. (2021). Deep Learning for Aspect-Based Sentiment Analysis on Indonesian Hotels Reviews. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *4*(3). https://doi.org/10.22219/kinetik.v6i3.1300

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176. https://doi.org/10.1145/363958.363994

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019*, *Mlm*, 4171–4186. https://aclanthology.org/N19-1423.pdf

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

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

Kokong, D. A. R. S., Irmawati, B., & Dwiyansaputra, R. (2024). Spelling Error Correction in Indonesian Using Damerau-Levenshtein Distance Dan N-Gram. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA)*, *6*(1), 257–263. https://doi.org/10.29303/jtika.v6i1.169

Nur, M. A. (2021). Perbandingan Levenshtein Distance Dan Jaro-Winkler Distance Untuk Koreksi Kata Dalam Preprocessing Analisis Sentimen Pengguna Twitter. *Jurnal Fokus Elektroda : Energi Listrik, Telekomunikasi, Komputer, Elektronika Dan Kendali)*, *6*(2), 88. https://doi.org/10.33772/jfe.v6i2.17751

Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 37–63. http://arxiv.org/abs/2010.16061

Pratama, A. Y., Sanjaya, G. A., Lubis, N. K., & Aditya, M. R. (2025). Analisis Sentimen Publik Terkait Danantara Menggunakan Algoritma IndoBERT pada Platform Media Sosial. *METIK Jurnal*, *9*(1), 2025. https://doi.org/10.47002/metik.v9i1.1055

Rahma, I. A., & Suadaa, L. H. (2023). Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(6), 1329–1340. https://doi.org/10.25126/jtiik.2023107325

Rahman, R. A., Pranatawijaya, V. H., & Sari, N. N. K. (2024). Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Gojek. *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, *4*(1), 70–82. https://doi.org/10.24002/konstelasi.v4i1.8922

Santoso, P., Yuliawati, P., Shalahuddin, R., & Wibawa, A. P. (2019). Damerau Levenshtein Distance for Indonesian Spelling Correction. *Jurnal Informatika*, *13*(2), 11. https://doi.org/10.26555/jifo.v13i2.a15698

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6382–6388. https://doi.org/10.18653/v1/d19-1670

Widansyah, M., Fathia Frazna Az-Zahra, & Agung Pambudi. (2024). Fine-Tuning Model Indobert (Indonesian Bidirectional Encoder Representations from Transformers) untuk Analisis Sentimen Berbasis Aspek pada Aplikasi M-Paspor. *Joutica*, *9*(2), 183–195. https://doi.org/10.30736/informatika.v9i2.1310

Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining