e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 9 No. 3, Desember, 2025, **Hal. 945-954**
DOI: 10.29408/edumatic.v9i3.32657

Benchmarking Model Transformer Modern untuk Analisis Sentimen dan Tren Konsumen dalam Industri Fashion

Daniel Fernando Herawan [1], Theresia Ratih Dewi Saputri [1, *]

[1] Program Studi Informatika, Universitas Ciputra Surabaya, Indonesia
[*] Correspondence: theresia.ratih@ciputra.ac.id

Copyright: © 2025 by the authors

Received: 15 Oktober 2025 | Revised: 2 November 2025 | Accepted: 4 Desember 2025 | Published: 12 Desember 2025

Abstrak
Dinamika tren dan preferensi konsumen dalam industri fashion menuntut analisis sentimen yang akurat untuk mendukung pengambilan keputusan. Penelitian ini bertujuan untuk membandingkan performa empat model *Transformer* yaitu BERT, DistilBERT, RoBERTa, dan DeBERTa dalam tugas klasifikasi sentimen ulasan produk *fashion* guna mengidentifikasi model yang paling efektif dalam memahami nuansa semantik yang khas dalam industri *fashion*. Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan 278.098 ulasan TeePublic yang diproses melalui pembersihan teks, penyeimbangan kelas, dan tokenisasi, kemudian dibagi menjadi data latih, validasi, dan uji. Empat model *Transformer* di-*fine-tuning* menggunakan *hyperparameter* hasil optimasi Optuna, dan dievaluasi menggunakan akurasi dan F1-score. Hasil temuan kami menunjukkan bahwa keempat model mencapai performa yang hampir setara, dengan F1-score dan akurasi berada pada rentang 0,824 - 0,826. Uji ANOVA turut mengonfirmasi tidak adanya perbedaan signifikan antar-model, sehingga peningkatan kompleksitas arsitektur tidak memberikan manfaat berarti pada tugas klasifikasi sentimen berskala terbatas. Temuan ini menunjukkan bahwa konsistensi kinerja ditentukan oleh kualitas pra-pemrosesan, strategi penyeimbangan kelas, dan desain *pipeline*. Penelitian ini menyediakan dasar komparatif untuk pemilihan model dalam analisis sentimen pada domain *fashion* dan menegaskan bahwa optimalisasi alur pemrosesan seringkali menghasilkan peningkatan yang lebih substansial dibandingkan beralih ke model dengan arsitektur yang lebih besar.

Kata kunci: analisis sentimen; *fast fashion*; *natural language processing*; *transformers*

Abstract
The dynamic trends and consumer preferences in the fashion industry demand accurate sentiment analysis to support decision-making. This study aims to compare the performance of four Transformer models BERT, DistilBERT, RoBERTa, and DeBERTa, in sentiment classification of fashion product reviews to identify which model best captures the semantic nuances specific to this domain. Using an experimental quantitative approach, we analyzed 278,098 TeePublic reviews that underwent text cleaning, class balancing, and tokenization, and were then split into training, validation, and test sets. The four models were fine-tuned using Optuna-optimized hyperparameters and evaluated with accuracy and F1-score. Our results show that all models achieved nearly identical performance, with F1-scores and accuracy ranging from 0.824 to 0.826. ANOVA testing confirmed no significant differences among models, indicating that increased architectural complexity does not yield substantial benefits for a task of this scale. These findings suggest that performance consistency is driven more by preprocessing quality, class-balancing strategies, and pipeline design. The study offers a comparative basis for model selection in fashion-domain sentiment analysis and highlights that optimizing the processing pipeline often brings more meaningful improvements than adopting larger architectures.

Keywords: *fast fashion; natural language processing; sentiment analysis; transformers*

Edumatic: Jurnal Pendidikan Informatika is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

PENDAHULUAN

Industri *fashion* merupakan salah satu sektor ekonomi terbesar dan paling cepat berubah di dunia (Kornelis, 2022). Industri ini mencakup berbagai tahapan seperti desain, produksi, distribusi, hingga pemasaran pakaian dan aksesori. Dalam beberapa tahun terakhir, munculnya model bisnis *fast fashion* telah mengubah cara produk *fashion* diproduksi dan dikonsumsi (Rufikasari, 2022). *Fast fashion* menekankan siklus produksi yang cepat, volume tinggi, dan biaya rendah untuk merespons tren pasar secara *real time* (Hakim & Rusadi, 2022). Namun, semakin tinggi peningkatan di industri ini juga menimbulkan konsekuensi serius, seperti peningkatan limbah tekstil, eksploitasi tenaga kerja, dan ketidakberlanjutan dalam rantai pasokan (Albab et al., 2024).

Salah satu tantangan utama dalam industri *fast fashion* adalah ketidaksesuaian antara penawaran dan permintaan, yang sering kali menyebabkan kelebihan produksi. Karena produksi yang berlebihan, beberapa perusahaan bahkan menghancurkan stok yang tidak terjual, dimana tindakan ini sangat tidak berkelanjutan (Ramadhan, 2024). Permasalahan ini berakar pada permintaan konsumen yang selalu berubah dengan cepat yang menyebabkan perusahaan untuk selalu memproduksi hal baru (Soni & Baldawa, 2023). Tidak seperti industri ritel lainnya yang bisa mengandalkan pola penjualan masa lalu, industri *fast fashion* sangat dipengaruhi oleh tren musiman dan preferensi konsumen yang berubah-ubah secara cepat, sehingga membuat prediksi permintaan menjadi kompleks dan penuh ketidakpastian (Rudniy et al., 2024). Kondisi ini menciptakan peluang baru, analisis sentimen terhadap ulasan produk dapat digunakan sebagai pendekatan untuk memahami tren *fashion* berbasis opini konsumen. Dengan menganalisis ekspresi dan persepsi konsumen terhadap produk *fashion*, perusahaan dapat memperoleh wawasan yang lebih akurat mengenai preferensi pasar dengan mudah dan cepat (Kadek & Diantari, 2021).

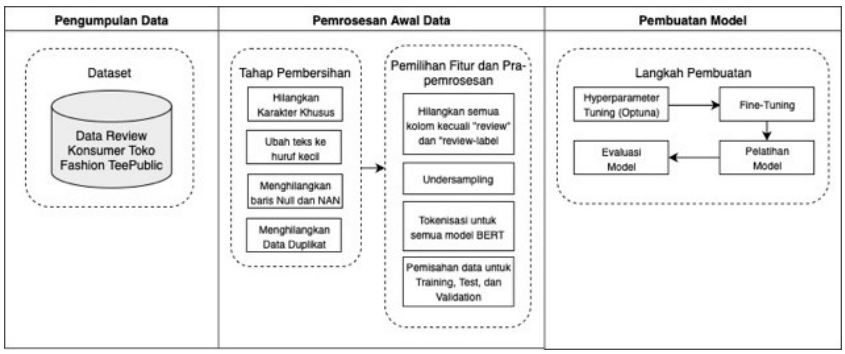
Pendekatan tradisional dalam peramalan tren *fashion* seperti *K-Nearest Neighbor* (KNN) dan *Linear Regression* memiliki keterbatasan dalam menangkap hubungan non-linear dan kompleks antara data penjualan (Ananta et al., 2025), sentimen pengguna (Rahayu, et al., 2022), preferensi konsumen, dan faktor eksternal (Efriadi et al., 2022; Kumar & Weissenberger-Eibl, 2024; Nabila et al., 2025; Pangestu et al., 2025). Beberapa tahun terakhir, perkembangan *Natural Language Processing* (NLP) menawarkan alternatif yang lebih adaptif untuk menganalisis opini konsumen dalam skala besar (Khadapi & Pakpahan, 2024). Salah satu kemajuan utama dalam NLP adalah model *transformer* dengan mekanisme *attention* yang memungkinkan pemahaman konteks bahasa secara lebih efisien. Model NLP konvensional seperti *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM) terbukti mampu mengenali pola sekuensial dalam teks, tetapi masih terbatas dalam menangkap dependensi jangka panjang dan konteks semantik yang kompleks, dimana model BERT mampu melakukannya (Berliana & Mustikasari, 2024).

Beberapa penelitian menunjukkan bahwa model pre-trained dapat memanfaatkan pengetahuan sebelumnya untuk beradaptasi dengan domain baru (Kim et al., 2023) dan evaluasi performanya biasanya dianalisis menggunakan metode statistik (Riaz et al., 2024) . Dari berbagai model NLP yang ada, *Bidirectional Encoder Representations from Transformers* (BERT) menunjukkan kinerja unggul dalam analisis sentimen dan telah diterapkan pada sektor *e-commerce*, industri film, sektor *game*, dan sosial media (Braja & Kodar, 2023; Karimah, 2024; Majid et al., 2023; Mangal & Makwana, 2025). Namun, belum banyak penelitian sebelumnya yang secara khusus mengevaluasi efektivitas model *transformer* pada domain *fashion*, yang memiliki karakteristik linguistik berbeda, misalnya penggunaan istilah deskriptif seperti “fit”, “color choice”, atau “eye catching” yang mencerminkan aspek visual dan emosional. Kondisi ini menimbulkan kesenjangan penelitian terkait kemampuan model memahami ekspresi linguistik khas *fashion*. Sehingga kebaruan dari penelitian ini terletak pada domain *fashion* yang memiliki linguistik khas.

Karakteristik linguistik yang emosional, subjektif, dan kontekstual dalam ulasan *fashion* menuntut model NLP yang adaptif. Penelitian ini bertujuan mengkaji komparasi kinerja empat arsitektur *transformer* yaitu BERT, DistilBERT, RoBERTa, dan DeBERTa dalam memodelkan opini konsumen dan memprediksi indikasi tren. Penelitian ini juga mengidentifikasi model yang paling kapabel dalam merepresentasikan nuansa semantik khas industri *fashion*. Penelitian ini memperluas pemahaman mengenai efektivitas model *transformer* dalam memproses teks dengan karakteristik domain-spesifik yang kompleks, serta menyediakan landasan teoretis bagi pengembangan sistem analitik yang mampu mengadaptasi dinamika linguistik dan preferensi pasar secara lebih responsif.

METODE

Penelitian ini mengadopsi pendekatan kuantitatif eksperimental dengan kerangka kerja yang digambarkan pada Gambar 1, yang memuat tiga tahapan utama: Pengumpulan Data, Pemrosesan Awal Data, dan Pembuatan Model. Tahap Pengumpulan Data fokus pada akuisisi ulasan konsumen dari *platform* TeePublic sebagai sumber data primer yang merepresentasikan opini pasar *fashion*. Pada tahap Pemrosesan Awal Data dilakukan serangkaian langkah pembersihan yaitu penghapusan karakter khusus, normalisasi huruf, penghilangan baris kosong dan duplikat. Dilanjutkan dengan pemilihan fitur sehingga hanya atribut "review" dan "review-label" yang dipertahankan. Setelah pembersihan, diterapkan strategi penyeimbangan kelas menggunakan *undersampling*, kemudian setiap teks ditokenisasi menggunakan *tokenizer* sesuai masing-masing model *transformer* dan *dataset* dibagi menjadi *subset training*, *validation*, dan *test*. Tahap Pembuatan Model meliputi optimasi *hyperparameter* menggunakan Optuna, proses *fine-tuning* model dengan Trainer API, serta evaluasi kinerja model menggunakan metrik klasifikasi.



Gambar 1. Diagram metodologi analisis tren fashion berbasis nlp

Dataset penelitian berasal dari platform Kaggle dan berisi ulasan konsumen terhadap produk *fashion* dari TeePublic. Dataset ini dipilih karena relevan dengan tujuan penelitian, yaitu memproyeksikan tren konsumen melalui analisis opini publik di industri fashion. Terdiri dari 278.098 baris data dengan dua atribut utama yaitu review (teks ulasan) dan review-label (skor sentimen 1–5). Skor 1–2 dipetakan sebagai sentimen negatif (0), skor 3 sebagai netral (1), dan skor 4-5 sebagai positif (2). Pengelompokan ini dilakukan karena 1-2 sama-sama merepresentasikan ketidakpuasan, sedangkan 4-5 menunjukkan kepuasan, dan aturan pemetaan ini dibuat eksplisit agar proses analisis dapat direproduksi. Tabel 1 menampilkan beberapa contoh representatif dari ulasan konsumen, termasuk label asli yang diberikan pada dataset serta label yang telah melalui proses konversi.

Eksplorasi awal menunjukkan adanya ketidakseimbangan label, dengan kelas positif sebagai mayoritas. Kondisi ini berpotensi menimbulkan bias, sehingga penelitian menerapkan teknik *undersampling* yang dinilai efektif mengurangi bias tanpa menambah kompleksitas serta cocok untuk data teks berdimensi tinggi (Jumansyah et al., 2024). Teks ulasan kemudian

ditokenisasi menggunakan *tokenizer* bawaan tiap model, dan *dataset* dibagi menjadi 80% pelatihan, 10% validasi, dan 10% pengujian untuk menjaga pemisahan evaluasi. Empat model yaitu BERT, DistilBERT, RoBERTa, dan DeBERTa dipilih karena perbedaan arsitekturalnya dapat memberikan perspektif performa yang lebih menyeluruh pada domain *fashion*. Seluruh model diimplementasikan menggunakan pustaka *Hugging Face Transformers* melalui *AutoModelForSequenceClassification.*

Tabel 1. Contoh ulasan konsumen

Review	Label Asli	Label Konversi
love the design, the quality is great	4	2
A bit pricey for a tee	1	1
2 out of the 4 T Shirts I received were the wrong color	3	0

Pelatihan model dilakukan menggunakan Trainer API dengan AdamW dan CrossEntropyLoss, serta *early stopping* untuk mencegah *overfitting*. Optimasi *hyperparameter* dilakukan menggunakan Optuna, dengan ruang pencarian berupa *learning rate* (1e-5 hingga 5e-5, log-scale), batch size {8, 16, 32}, jumlah *epoch* (2-5), dan *weight decay* (0.0-0.3). Optuna memaksimalkan macro F1-score pada *validation set* sebagai *objective function,* sementara akurasi tetap digunakan sebagai metrik pelengkap. Macro F1 dipilih karena lebih representatif pada *dataset* yang awalnya tidak seimbang. Untuk mengevaluasi apakah perbedaan performa antar model signifikan secara statistik, penelitian ini menggunakan One-Way ANOVA (Wang & Chen, 2025). Asumsi ANOVA telah diperiksa dan tidak menunjukkan pelanggaran berarti, sehingga analisis dapat dilakukan. Dalam konteks penelitian ini, ANOVA berfungsi sebagai pemeriksa kesenjangan performa, bukan penentu model terbaik.

HASIL DAN PEMBAHASAN

Hasil

Eksplorasi awal dataset menunjukkan adanya ketidakseimbangan kelas yang cukup besar, dengan 203.293 ulasan positif, 27.444 netral, dan 15.728 negatif dari total 278.098 data. Ketimpangan ini berpotensi menyebabkan model terlalu bergantung pada pola kelas mayoritas. Setelah dilakukan *random undersampling* pada kelas Positif hingga berjumlah 27.444 sampel, distribusi antar kelas menjadi lebih seimbang dengan komposisi data positif sebanyak 27.444, data netral sebanyak 27.444, dan data negatif sebanyak 15.728. Perubahan ini menghasilkan *set* data latih yang lebih proporsional untuk mengevaluasi performa model secara adil. Setelah proses optimasi menggunakan Optuna, setiap model memperoleh konfigurasi *hyperparameter* terbaik yang disesuaikan dengan karakteristik arsitektur masing-masing. Optimasi dilakukan terhadap empat parameter utama, yaitu *learning rate*, *batch size*, jumlah *epoch*, dan *weight decay*.

Tabel 2. Nilai *hyperparameter* terbaik tiap model

Model	*Learning Rate*	*Batch Size*	Jumlah Epoch	*Weight Decay*
BERT	1,06e-05	16	4	0,16
DistilBERT	3,42e-05	32	2	0,2
RoBERTa	1,56e-05	8	2	0,26
DeBERTa	1,16e-05	16	2	0,28

Setelah proses pelatihan, dilakukan perbandingan kinerja empat model tersebut yang telah melalui fine-tuning menggunakan konfigurasi *hyperparameter* optimal hasil optimasi dengan Optuna. Tabel 2 menunjukkan bahwa setiap model memerlukan konfigurasi

hyperparameter yang berbeda untuk mencapai performa terbaik, yang mengindikasikan adanya sensitivitas arsitektur terhadap pengaturan parameter pelatihan. Sebagai contoh, DistilBERT mencapai performa optimal pada *batch size* yang lebih besar dibandingkan model lainnya, sedangkan BERT membutuhkan jumlah *epoch* lebih tinggi untuk mendapatkan hasil terbaik, mencerminkan kebutuhan komputasi yang lebih besar. Sementara itu, variasi nilai l*earning rate* dan *weight decay* pada masing-masing model menunjukkan perbedaan stabilitas selama proses *fine-tuning*. Evaluasi performa model kemudian dilakukan pada dua kondisi, yaitu sebelum dan sesudah proses *undersampling*, untuk mengamati bagaimana distribusi kelas memengaruhi kualitas klasifikasi sentimen.

Tabel 3. Hasil evaluasi model sebelum *undersampling*

Model	BERT	DistilBERT	RoBERTa	DeBERTa
Akurasi	0,927	0,925	0,927	0,931
F1-Score	0,925	0,923	0,926	0,931
Waktu Latih (Menit)	113	34	92	105

Hasil evaluasi model sebelum dilakukan *undersampling* yang disajikan pada tabel 3 menujukkan bahwa seluruh model menunjukkan performa yang tinggi dengan akurasi dan F1-score berada pada kisaran 0,92. Pola ini mengindikasikan bahwa keempat arsitektur *transformer* mampu menangkap karakteristik linguistik domain *fashion* dengan baik ketika jumlah data latih masih penuh. Meskipun terdapat variasi kecil antar model, performanya relatif seragam, mencerminkan konsistensi representasi semantik yang dihasilkan oleh model-model tersebut dalam kondisi distribusi data yang tidak seimbang.

Tabel 4. Hasil evaluasi model sesudah *undersampling*

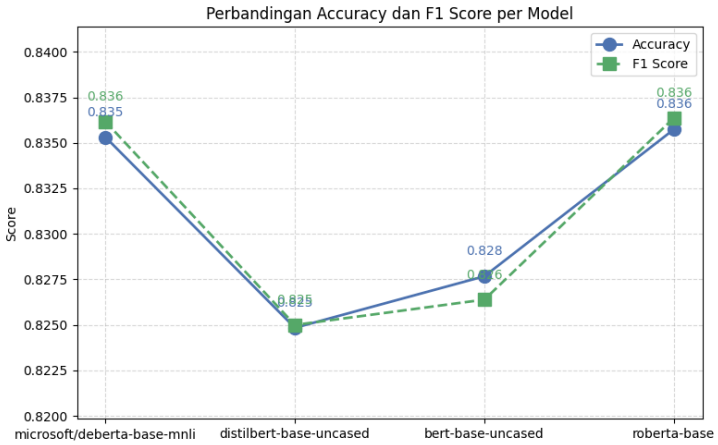
Model	BERT	DistilBERT	RoBERTa	DeBERTa
Akurasi	0,827	0,824	0,836	0,836
F1-Score	0,826	0,824	0,836	0,835
Waktu Latih (Menit)	32	10	26	30

Selanjutnya hasil evaluasi model setelah dilakukan *undersampling* pada tabel 4 menujukkan bahwa nilai akurasi dan F1-score menurun ke kisaran 0,82 pada seluruh model. Penurunan skor ini tidak terkait dengan perubahan distribusi semantik antar kelas, karena proporsi leksikal dan struktur kalimat tetap terjaga, yang berkurang adalah variasi data akibat pengurangan jumlah contoh latih pada kelas mayoritas. Meskipun performanya lebih rendah dibandingkan kondisi awal, hasil ini memberikan estimasi kinerja yang lebih tidak bias dan lebih representatif, terutama karena ketidakseimbangan kelas yang sebelumnya dominan telah diperbaiki melalui proses *undersampling*.

Efisiensi pelatihan menunjukkan bahwa DistilBERT memiliki waktu latih tercepat, diikuti oleh RoBERTa dan DeBERTa, sedangkan BERT memerlukan durasi pelatihan paling lama. Perbedaan ini mengindikasikan adanya *gap* efisiensi komputasi yang signifikan meskipun selisih akurasi antarmodel relatif kecil. Waktu pelatihan BERT yang lebih panjang juga tidak menghasilkan peningkatan performa yang sebanding, karena akurasi dan F1-score-nya tetap berada di bawah RoBERTa dan DeBERTa. Sebaliknya, kedua model tersebut memberikan keseimbangan yang lebih baik antara durasi pelatihan dan kinerja akhir, mencerminkan efisiensi arsitektur yang lebih modern. Pola tersebut menegaskan adanya *trade-off* kuantitatif antara ukuran model dan kebutuhan komputasi. Visualisasi perbandingan akurasi dan F1-score dapat dilihat pada gambar 2.

Penelitian ini menerapkan uji statistik One-Way ANOVA pada nilai prediksi numerik yang dihasilkan oleh masing-masing model untuk menilai apakah perbedaan performa yang

muncul bersifat signifikan secara statistik atau hanya terjadi secara acak. Pendekatan ini penting karena ANOVA memungkinkan evaluasi komparatif lebih mendalam terhadap variasi antar model, khususnya ketika keempat arsitektur memiliki skor akurasi dan F1 yang relatif berdekatan setelah proses undersampling. Hasil dari pengujian ANOVA disajikan pada Tabel 5.



Gambar 2. Hasil perbandingan evaluasi akurasi dan f1-score model

Tabel 5. Hasil pengujian statistik anova

	F-statistic	***p-value***
Nilai	1,2573	2,8721e-01

Berdasarkan hasil pengujian ANOVA pada tabel 5, bahwa tidak terdapat perbedaan signifikan antar model *transformer* dalam melakukan klasifikasi sentimen pada dataset yang sama. Dengan kata lain, performa model setara secara statistik. Hasil ini tidak menghapus perbedaan praktis yang terlihat dari akurasi dan Macro F1-score, tetapi menegaskan bahwa variasi performa antar model relatif stabil. Dengan demikian, ANOVA berfungsi sebagai pemeriksaan konsistensi global, sementara evaluasi utama tetap berpusat pada metrik praktis yang menunjukkan keunggulan model-model *transformer* modern pada domain *fashion* yang bersifat kontekstual, sehingga penelitian ini turut mengisi celah penelitian sebelumnya.

Pembahasan

Distribusi akhir yang dihasilkan melalui *undersampling* tidak sepenuhnya seimbang secara numerik, karena strategi ini dilakukan secara proporsional untuk mempertahankan keberagaman konteks ulasan. Pendekatan tersebut bertujuan menghindari hilangnya informasi penting dari kelas mayoritas, yang sering kali mengandung variasi ekspresi positif yang kaya dan relevan terhadap konteks industri *fashion*. Strategi ini memungkinkan model untuk tetap mempelajari distribusi emosi dan konteks yang beragam tanpa kehilangan kemampuan generalisasi. BERT berfungsi sebagai *baseline transformer* penuh dan menunjukkan konfigurasi optimal pada *learning rate* kecil (1,06e-05) karena arsitektur berparameter besar membutuhkan pembaruan yang lebih hati-hati untuk mencegah *overfitting*. DistilBERT, dengan struktur yang lebih ringan dan lebih dangkal, mencapai performa terbaik menggunakan *batch size* lebih besar (32) dan *epoch* lebih sedikit (2) karena model ini cepat konvergen dan tidak memerlukan update berulang untuk menstabilkan representasinya. RoBERTa menunjukkan kombinasi optimal berupa *learning rate* menengah (1,56e-05) dan *weight decay* tinggi (0,26), yang sesuai dengan arsitekturnya yang dilatih tanpa NSP dan cenderung menghasilkan gradien lebih besar sehingga membutuhkan regularisasi kuat. DeBERTa

menemukan konfigurasi optimal dengan *learning rate* rendah (1,16e-05) dan *weight decay* besar (0,28), selaras dengan mekanisme *disentangled attention* yang sensitif terhadap pembaruan agresif dan membutuhkan langkah belajar halus agar relasi semantik yang dipisahkan tetap stabil. Secara keseluruhan, pola ini menegaskan bahwa setiap arsitektur memiliki kebutuhan pelatihan intrinsik yang berbeda, sehingga penggunaan *hyperparameter* optimal untuk masing-masing model memastikan perbandingan performa pada tahap evaluasi tetap *fair*.

Temuan ini juga konsisten dengan teori saturasi performa dalam NLP, yaitu ketika model-model dengan kompleksitas berbeda menghasilkan akurasi yang serupa karena seluruh arsitektur telah cukup mampu menangkap pola linguistik utama dalam dataset yang homogen. Dari sisi praktis, hasil uji ANOVA memberikan implikasi penting. Pemilihan model *transformer* dapat disesuaikan dengan kebutuhan efisiensi, karena performa antar model tidak berbeda signifikan secara statistic dan sejalan dengan penelitian Aliyu et al. (2024) yang menemukan perbedaan akurasi atau F1-score antar model relatif kecil sehingga pemilihan model dapat mempertimbangkan efisiensi dan ketersediaan sumber daya komputasi. DistilBERT, misalnya, dapat diprioritaskan untuk sistem yang membutuhkan efisiensi komputasi tanpa kehilangan akurasi yang berarti. Ketidaksignifikanan ANOVA bukan hanya menandakan kesetaraan performa, ini juga mengisyaratkan bahwa ketika fondasi metodologis diterapkan dengan baik, bahkan model yang lebih ringan dapat mencapai performa sekelas model besar. Temuan ini menantang asumsi umum bahwa peningkatan kompleksitas model selalu menghasilkan peningkatan akurasi, dan justru menegaskan bahwa desain *pipeline* yang kokoh memiliki kontribusi yang lebih besar terhadap hasil akhir. Dengan demikian, ANOVA dalam penelitian ini tidak hanya berfungsi sebagai verifikasi statistik, tetapi juga menjadi bukti bahwa optimalisasi proses dan kualitas data memiliki peran yang lebih krusial dibandingkan sekadar memilih arsitektur *transformer* yang lebih canggih.

Sebagai titik rujukan terhadap temuan-temuan dalam penelitian sebelumnya, hasil penelitian ini menunjukkan pola yang sejalan sekaligus berbeda secara kontekstual. Penelitian Joshy & Sundar (2022) melaporkan bahwa BERT mengungguli DistilBERT dan RoBERTa pada dataset *tweet* yang bersifat generik dan berbahasa pendek, sedangkan dalam domain *fashion* pada penelitian ini, RoBERTa dan DeBERTa justru menunjukkan performa lebih stabil. Perbedaan ini menguatkan bahwa efektivitas arsitektur *transformer* sangat dipengaruhi oleh karakteristik domain dan struktur linguistik teks. Selain itu, penelitian Wang & Chen (2025) yang menyoroti pentingnya *domain adaptation* dalam analisis dialog pendidikan turut mendukung temuan ini, bahwa performa model bukan hanya ditentukan oleh kompleksitas arsitektur, tetapi oleh kesesuaian antara representasi pra-latih dan sifat domain target. Dengan demikian, pembandingan ini memperkuat interpretasi bahwa pemilihan model dalam penelitian *sentiment analysis* tidak dapat dilepaskan dari konteks domain, dan bahwa seluruh model dalam penelitian ini bekerja secara fungsional setara justru karena metodologi *balancing* dan pra-pemrosesan yang solid berhasil meminimalkan bias antar arsitektur.

Berdasarkan seluruh temuan empiris, penelitian ini juga menyajikan kerangka pemilihan model yang lebih aplikatif bagi berbagai kebutuhan penggunaan. RoBERTa dan DeBERTa direkomendasikan ketika konteks linguistik kaya nuansa, ekspresif, dan domain-spesifik yaitu aspek yang umum pada ulasan *fashion* menjadi prioritas utama, karena keduanya menunjukkan kapasitas representasi semantik yang paling kuat. Untuk skenario dengan keterbatasan sumber daya komputasi atau kebutuhan inferensi cepat, DistilBERT menjadi pilihan yang paling efisien tanpa penurunan akurasi yang berarti. Sementara itu, BERT tetap relevan sebagai *baseline* yang stabil ketika diperlukan model standar dengan perilaku yang dapat diprediksi. Dengan demikian, temuan ini tidak hanya membandingkan performa model, tetapi juga memberikan panduan yang jelas mengenai kondisi penggunaan optimal bagi masing-masing arsitektur.

Penelitian ini memiliki beberapa keterbatasan, antara lain evaluasi model yang belum menggunakan *cross-validation* sehingga performa mungkin kurang stabil, serta tidak dilakukan *error analysis* sehingga jenis kesalahan model belum teridentifikasi. Untuk penelitian selanjutnya, evaluasi dapat diperluas dengan metrik ROC-AUC serta analisis statistik tambahan, seperti post hoc Tukey HSD, post hoc Nemenyi, atau uji Friedman, untuk membandingkan performa model secara lebih komprehensif dan meningkatkan interpretabilitas hasil.

SIMPULAN

Penelitian ini menunjukkan bahwa model *transformer* modern, seperti DeBERTa dan RoBERTa, mampu menangkap pola linguistik emosional dan kontekstual dalam ulasan produk *fashion*, yang dikenal memiliki ekspresi bahasa khas dan nuansa subjektif. Temuan ini selaras dengan teori NLP yang menyatakan bahwa arsitektur *transformer* dengan mekanisme *attention* dapat memodelkan dependensi semantik yang kompleks dalam teks panjang. Dari perspektif praktis, pemahaman ini memberikan dasar bagi pengembangan sistem deteksi tren konsumen yang lebih adaptif, memungkinkan perusahaan *fashion* untuk memantau preferensi pasar secara *real-time* dan merespons dinamika tren dengan cepat. Meski performa model relatif seimbang menurut uji statistik, penelitian ini memiliki beberapa keterbatasan, seperti penggunaan *undersampling* yang mengurangi variasi data mayoritas dan evaluasi tanpa *cross-validation*, serta minimnya analisis kesalahan spesifik. Penelitian selanjutnya dapat mengeksplorasi teknik *balancing* alternatif seperti SMOTE atau *ensemble balancing,* integrasi *domain adaptation* untuk menangani pergeseran distribusi bahasa dalam ulasan *fashion*, serta penerapan metrik dan uji statistik tambahan seperti ROC-AUC, uji Friedman, dan *post hoc* Tukey HSD atau Nemenyi, pengujian menggunakan *dataset* lain juga pilihan yang layak. Pendekatan ini tidak hanya akan meningkatkan robustitas evaluasi model, tetapi juga memperdalam pemahaman tentang bagaimana karakteristik linguistik domain *fashion* memengaruhi performa model *transformer*.

REFERENSI

Albab, W. U., Mardiah, A. R., Ranjani, G., Karina, G. D., & Safitri, M. N. (2024). Pengaruh Industri Fast Fashion Terhadap Pencemaran Lingkungan dan Penurunan Keadilan Antar Generasi. *Indonesian Journal of Criminal Law and Criminology*, *5*(3), 94–103. https://doi.org/https://doi.org/10.18196/ijclc.v5i3.22830

Aliyu, Y., Sarlan, A., Danyaro, K. U., Sani, A., & Rahman, B. A. (2024). Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages. *International Journal of Advanced Computer Science and Applications*, *15*(4). https://doi.org/10.14569/IJACSA.2024.0150437

Ananta, A. Y., Ariyanto, R., Rozi, I. F., & Arianto, R. (2025). Analisis Performa Metode Extreme Learning Machine dan Multiple Linear Regression dalam Prediksi Produksi Gula. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 169–178. https://doi.org/10.29408/edumatic.v9i1.29626

Berliana, A. S., & Mustikasari, M. (2024). Analisis Sentimen pada Ulasan Aplikasi JakartaNotebook di Google Play Menggunakan Metode Recurrent Neural Network (RNN). *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3), 2830–7062. https://doi.org/10.23960/jitet.v12i3.5067

Braja, A. S. P., & Kodar, A. (2023). Implementasi Fine-Tuning BERT untuk Analisis Sentimen terhadap Review Aplikasi PUBG Mobile di Google Play Store. *Jurnal Informatika Merdeka Pasuruan*, *7*(3), 120. https://doi.org/10.51213/jimp.v7i3.779

Efriadi, D., Rahmaddeni, R., Agustin, A., & Junadhi, J. (2022). Prediksi Penambahan Piutang Iuran Jaminan Sosial Ketenagakerjaan menggunakan Algoritma K-Nearest

Neighbor. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 49–57. https://doi.org/10.29408/edumatic.v6i1.5255

Hakim, A. L., & Rusadi, E. Y. (2022). Kritik Globalisasi: Fenomena Fast Fashion Sebagai Budaya Konsumerisme Pada Kalangan Pemuda Kota Surabaya. ALMAARIEF, 59-67. https://doi.org/10.35905/almaarief.v4i2.2768

Joshy, A., & Sundar, S. (2022). Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa. *International Power and Renewable Energy Conference (IPRECON)*, 1-6. India: IEEE https://doi.org/10.1109/IPRECON55716.2022.10059542

Jumansyah, L., Mohamad Soleh, A., Dyah Syafitri, U., Risman Dwi, L., Mohamad, A., & Dyah, U. (2024). Manifold Learning and Undersampling Approaches for Imbalanced Class Sentiment Classification. *Knowledge Engineering and Data Science*, *7*(2), 139–151.

Kadek, N., & Diantari, Y. (2021). Trend Cycle Analysis on Fast Fashion Products. *Journal of Aesthetics, Design, and Art Management, 1*(1), 24–33. https://doi.org/10.58982/jadam.v1i1.101

Karimah, N. (2024). Multi-Aspect Sentiment Analysis Pada Review Film Menggunakan Metode Bidirectional Encoder Representations from Transformers (BERT). *Komputika : Jurnal Sistem Komputer*, *13*(1), 63–72. https://doi.org/10.34010/komputika.v13i1.11098

Khadapi, M., & Pakpahan, V. M. (2024). Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024. *Jurnal Komputer dan Informatika*. Jurnal Komputer dan Informatika, 6(2), 130–137.

Kim, H. J., Cho, H., Lee, S. W., Kim, J., Park, C., Lee, S. G., Yoo, K. M., & Kim, T. (2023). Universal Domain Adaptation for Robust Handling of Distributional Shifts in NLP. *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2023*, 1(1), 5888–5905. https://doi.org/10.18653/v1/2023.findings-emnlp.392

Kornelis, Y. (2022). Fenomena Industri Fast Fashion: Kajian Hukum Perspektif Kekayaan Intelektual Indonesia. *Jurnal Komunitas Yustisia*, *5*(1), 262–277. https://doi.org/10.23887/jatayu.v5i1.46040

Kumar, D., & Weissenberger-Eibl, M. (2024, June 19). Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights. *R&D Management Conference 2024*. Stockholm: Fraunhover. https://doi.org/10.24406/publica-3456

Majid, A., Nugraha, D., & Adhinata, F. D. (2023). Sentiment Analysis on Tiktok Application Reviews Using Natural Language Processing Approach. *Journal of Embedded Systems, Security and Intelligent Systems*, *4*(1), 32–38. https://doi.org/10.59562/jessi.v4i1.471

Mangal, D., & Makwana, H. (2025). Performance Analysis of Different BERT Implementation for Event Burst Detection from Social Media Text. *Indonesian Journal of Electrical Engineering and Computer Science*, *38*(1), 439–446. https://doi.org/10.11591/ijeecs.v38.i1.pp439-446

Nabila, P., Wajidi, F., & Firgiawan, W. (2025). Analisis Algoritma K-Nearest Neighbor dan Naive Bayes pada Kebijakan Perkuliahan Online dengan Multi Bahasa. *Techno.Com*, *24*(2), 524–542. https://doi.org/10.62411/tc.v24i2.12656

Pangestu, A. G., Winarno, S., Nugraha, A., & Muttaqin, A. N. I. (2025). DiabTrack: Sistem Prediksi Dini Diabetes Melitus Tipe 2 berbasis Web menggunakan Algoritma K-Nearest Neighbors. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 284–293. https://doi.org/10.29408/edumatic.v9i1.29691

Rahayu, S., MZ, Y., Bororing, J. E., & Hadiyat, R. (2022). Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi

Finansial FLIP. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 98–106. https://doi.org/10.29408/edumatic.v6i1.5433

Ramadhan, A. F. (2024). Tren Fast Fashion Pakaian Masa New Normal di Indonesia: Efektivitas Konsep Sustainable Fashion Terhadap Lingkungan. *Journal of Waste and Sustainable Consumption*, *1*(2), 77–89. https://doi.org/10.61511/jwsc.v1i2.2024.1247

Riaz, N., Mehboob, H., & Wajid, S. (2024). Natural Language Processing For Sentiment Analysis: A Machine Learning Approach. *Contemporary Journal of Social Science Review*, *2*(04), 106-116.

Rudniy, A., Rudna, O., & Park, A. (2024). Trend tracking tools for the fashion industry: the impact of social media. *Journal of Fashion Marketing and Management*, *28*(3), 503–524. https://doi.org/10.1108/JFMM-08-2023-0215

Rufikasari, Y. D. (2022). Telaah Teologi, Ekonomi dan Ekologi Terhadap Fenomena Fast Fashion Industry. *Teologis, Relevan, Aplikatif, Cendikia, Kontekstual*, *1*(2), 64–83. https://doi.org/10.61660/tep.v1i2.23

Soni, S., & Baldawa, S. (2023). Analyzing Sustainable Practices in Fashion Supply Chain. *International Research Journal of Business Studies*, *16*(1), 11–25. https://doi.org/10.21632/irjbs.16.1.11-25

Wang, D., & Chen, G. (2025). Evaluating the Use of BERT and Llama to Analyse Classroom Dialogue for Teachers' Learning of Dialogic Pedagogy. *British Journal of Educational Technology*, *56(6)*, 2671–2704. https://doi.org/10.1111/bjet.13604