

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi sudah mengubah cara orang berbelanja. Sekarang perdagangan *online* atau *e-commerce* menjadi bagian penting dalam kegiatan ekonomi masyarakat Indonesia. Platform seperti tokopedia, shopee, dan bukalapak bukan hanya tempat belanja, tapi juga ruang di mana pembeli dan penjual bisa berinteraksi lewat fitur ulasan produk. Setiap hari, ribuan orang membagikan ulasan mereka yang menunjukkan pengalaman, kepuasan, dan perasaan terhadap produk yang sudah dibeli (Aryani & Wandebori, 2023; Hidayat *et al*, 2021). Di bidang kecantikan, ulasan produk sangat penting sebagai salah satu faktor yang mempengaruhi orang memutuskan untuk membeli, terutama produk perawatan wajah seperti serum yang membutuhkan kepercayaan tinggi.

Saat ini, industri kecantikan Indonesia diwarnai oleh munculnya berbagai merek lokal yang mampu bersaing dengan produk dari luar negeri. Beberapa merek seperti Somethinc, Skintific, Avoskin, dan MS Glow telah mendapat popularitas tinggi di kalangan konsumen Indonesia (Ramadhani & Purnomo, 2022). Di antara berbagai kategori produk yang dikembangkan, serum wajah menjadi salah satu kategori yang paling banyak diminati, didorong oleh meningkatnya kebutuhan masyarakat akan produk perawatan kulit yang efektif namun terjangkau (Putri *et al*, 2023).

Serum wajah adalah produk perawatan kulit yang memiliki fungsi lebih khusus dibandingkan produk perawatan dasar, contohnya untuk mencerahkan, melembapkan, atau mengatasi masalah jerawat (Putri *et al*, 2023). Karena efektivitasnya sangat tergantung pada kesesuaian kulit setiap individu, calon pembeli biasanya memperhatikan pengalaman pengguna lain melalui ulasan sebelum memutuskan untuk membeli (Aryani & Wandebori, 2023). Oleh karena itu, penelitian ini memanfaatkan data ulasan dari sepuluh merek serum *skincare*

lokal yang terkenal di platform Tokopedia sebagai objek analisis sentimen. Pemilihan merek dilakukan berdasarkan popularitas produk dan ketersediaan ulasan yang cukup, sehingga diharapkan dataset yang diperoleh dapat mencerminkan pola sentimen konsumen terhadap produk serum lokal dengan lebih utuh

Namun, jumlah ulasan yang sangat banyak di platform *e-commerce* memberikan tantangan bagi berbagai pihak. Di satu sisi, konsumen kesulitan memahami keseluruhan perasaan dari ratusan hingga ribuan ulasan yang tersedia untuk satu produk. Di sisi lain, produsen membutuhkan cara yang efisien untuk menganalisis tanggapan konsumen secara keseluruhan agar bisa mendukung pengambilan keputusan strategis (Zhang *et al*, 2021). Dalam situasi seperti ini, analisis sentimen menjadi solusi yang tepat untuk mengklasifikasikan pendapat konsumen secara otomatis ke dalam kategori positif, netral, atau negatif, dengan penerapan model pra-latih berbasis Transformer yang telah terbukti efektif pada analisis sentimen berbahasa Indonesia (Birjali *et al*, 2021; Ahmadian *et al*, 2024).

Salah satu tantangan utama dalam pengembangan sistem analisis sentimen adalah distribusi data yang tidak seimbang (*class imbalance*) antar kelas sentimen. Pada data ulasan produk di platform *e-commerce*, ulasan dengan sentimen positif umumnya jauh lebih dominan dibandingkan ulasan netral maupun negatif. Kondisi tersebut dapat menyebabkan model pembelajaran mesin cenderung lebih baik dalam mengenali kelas mayoritas, namun mengalami penurunan kinerja ketika mengklasifikasikan kelas minoritas. Selain itu, ulasan konsumen dalam Bahasa Indonesia juga banyak menggunakan bahasa tidak formal, singkatan, kesalahan penulisan, serta emoji yang dapat meningkatkan kompleksitas proses analisis sentimen (Kuriawan *et al.*, 2021). Oleh karena itu, diperlukan pendekatan yang mampu memperkaya variasi data pelatihan sehingga model dapat mempelajari karakteristik setiap kelas sentimen secara lebih seimbang.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik augmentasi data sebagai salah satu strategi untuk memperkaya variasi data

pelatihan dan membantu mengurangi ketidakseimbangan distribusi kelas. Data *augmentation* merupakan proses menghasilkan variasi baru dari data yang telah ada tanpa mengubah makna atau label sentimen aslinya sehingga model dapat mempelajari pola bahasa yang lebih beragam dan meningkatkan kemampuan generalisasi selama proses pelatihan (Natasya & Girsang, 2023). Dalam penelitian ini digunakan dua teknik augmentasi data, yaitu *back-translation* dan *synonym replacement*. Kedua teknik tersebut dipilih karena banyak digunakan untuk menangani data teks yang tidak seimbang dan mampu meningkatkan keragaman data pelatihan sehingga membantu meningkatkan performa model klasifikasi teks berbahasa Indonesia (Ifitah & Lya, 2023).

Selain penerapan teknik augmentasi data, penelitian ini juga menggunakan model IndoBERT, yaitu model bahasa berbasis Transformer yang telah dilatih sebelumnya menggunakan korpus teks berbahasa Indonesia dalam jumlah besar. IndoBERT dipilih karena mampu memahami karakteristik bahasa Indonesia, termasuk penggunaan bahasa tidak baku yang umum ditemukan pada ulasan produk di platform e-commerce. Selanjutnya, model tersebut di-*fine-tune* menggunakan data ulasan serum *skincare* lokal yang telah melalui tahap augmentasi sehingga dapat mengklasifikasikan sentimen ke dalam kelas positif, netral, dan negatif secara lebih akurat.

Berbagai penelitian sebelumnya telah menerapkan analisis sentimen menggunakan model berbasis Transformer maupun IndoBERT pada berbagai domain, seperti media sosial dan ulasan produk secara umum (Susanti *et al*, 2022; Sutoyo & Chowanda, 2021). Namun, penelitian yang secara khusus mengintegrasikan teknik *back-translation* dan *synonym replacement* pada proses *fine-tuning* IndoBERT untuk analisis sentimen ulasan serum *skincare* lokal di platform Tokopedia masih sangat terbatas. Oleh karena itu, penelitian ini memfokuskan analisis pada ulasan produk serum *skincare* lokal di platform Tokopedia. Selain itu, penelitian ini mengintegrasikan teknik *data augmentation* menggunakan *back-translation* dan *synonym replacement* dengan proses *fine-tuning* IndoBERT untuk mengatasi ketidakseimbangan distribusi kelas pada data pelatihan. Pendekatan tersebut diharapkan mampu menghasilkan model yang

lebih baik dalam mengklasifikasikan sentimen ulasan berbahasa Indonesia.

Meskipun berbagai penelitian terdahulu telah menunjukkan bahwa model berbasis IndoBERT mampu menghasilkan performa yang baik dalam analisis sentimen berbahasa Indonesia, masih terdapat beberapa keterbatasan yang perlu dikembangkan. Namun demikian, masih terdapat beberapa keterbatasan yang perlu dikembangkan. Sebagian besar penelitian menggunakan *dataset* umum sehingga belum secara khusus merepresentasikan karakteristik bahasa pada ulasan produk *skincare* lokal. Selain itu, permasalahan ketidakseimbangan distribusi kelas sentimen masih menjadi tantangan yang dapat memengaruhi kemampuan model dalam mengenali kelas minoritas. Penelitian yang menggabungkan proses *fine-tuning* IndoBERT dengan teknik augmentasi data pada domain ulasan *skincare* lokal serta mengimplementasikannya ke dalam aplikasi web juga masih relatif terbatas. Oleh karena itu, penelitian ini dilakukan untuk mengatasi keterbatasan tersebut melalui penerapan *fine-tuned* IndoBERT dengan teknik augmentasi data pada ulasan produk *skincare* lokal serta mengimplementasikan hasilnya dalam bentuk aplikasi web analisis sentimen.

Kebaruan (*novelty*) penelitian ini terletak pada penerapan model *Fine-tuned* IndoBERT yang dikombinasikan dengan teknik augmentasi data menggunakan *back-translation* dan *synonym replacement* yang diterapkan pada *dataset* ulasan sepuluh merek serum *skincare* lokal Indonesia. Selain itu, penelitian ini tidak hanya berfokus pada pengembangan model klasifikasi, tetapi juga mengimplementasikan model hasil pelatihan ke dalam aplikasi web analisis sentimen sehingga dapat digunakan secara langsung untuk membantu proses analisis ulasan produk.

Berdasarkan uraian tersebut, penelitian ini mengembangkan sistem analisis sentimen ulasan produk serum *skincare* lokal di platform Tokopedia menggunakan model *fine-tuned* IndoBERT yang didukung teknik augmentasi data berupa *back-translation* dan *synonym replacement*. Penerapan augmentasi data diharapkan dapat meningkatkan variasi data latih pada kelas minoritas sehingga membantu mengurangi ketidakseimbangan distribusi kelas dan meningkatkan kemampuan model dalam mengklasifikasikan sentimen ulasan.

Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem analisis sentimen berbahasa Indonesia, khususnya pada domain ulasan produk *skincare* lokal.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah yang akan dibahas dalam penelitian ini adalah:

1. Bagaimana membangun model analisis sentimen menggunakan *fine-tuned* IndoBERT untuk ulasan serum *skincare* lokal?
2. Bagaimana penerapan teknik augmentasi data dalam meningkatkan kualitas data pelatihan pada proses fine-tuning IndoBERT?
3. Bagaimana membangun aplikasi web yang dapat menganalisis sentimen ulasan produk serum secara otomatis?

1.3 Tujuan Penelitian

Adapun tujuan yang ingin dicapai pada penelitian ini adalah:

1. Membangun model analisis sentimen berbasis *fine-tuned* IndoBERT untuk mengklasifikasikan sentimen ulasan produk serum *skincare* lokal di tokopedia.
2. Menerapkan teknik augmentasi data (*back-translation* dan *synonym replacement*) pada data pelatihan untuk mendukung proses *fine-tuning* model IndoBERT.
3. Membangun aplikasi web berbasis *sentiment analysis* yang dapat digunakan untuk menganalisis sentimen ulasan produk secara real-time.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Penelitian ini diharapkan memberikan kontribusi bagi pengembangan teknologi pemrosesan bahasa alami, khususnya dalam penerapan *fine-tuned* IndoBERT dengan augmentasi data untuk analisis sentimen berbahasa Indonesia dengan *dataset* terbatas.
2. Brand *skincare* lokal dan platform *e-commerce* dapat mengaplikasikan

model ini untuk menganalisis sentimen ulasan konsumen secara otomatis, sehingga mendukung pengambilan keputusan strategis dan meningkatkan kualitas layanan berbasis *feedback* konsumen.

3. Calon konsumen dapat memperoleh pemahaman yang lebih cepat mengenai sentimen keseluruhan terhadap produk serum *skincare* lokal tanpa harus membaca ratusan ulasan satu per satu, sehingga meningkatkan efisiensi dalam proses pengambilan keputusan pembelian.
4. Penelitian ini dapat menjadi dasar bagi pengembangan dan penerapan model Transformer dengan teknik augmentasi data pada domain *e-commerce* lainnya atau *task* NLP berbahasa Indonesia dengan data terbatas pada masa mendatang.

1.5 Ruang Lingkup dan Batasan Penelitian

Adapun Ruang Lingkup dan Batasan Penelitian ini agar tujuan tercapai maka:

1. Penelitian ini hanya difokuskan pada analisis sentimen ulasan produk serum *skincare* dari brand lokal Indonesia di platform tokopedia dan tidak mencakup platform *e-commerce* lainnya.
2. Data yang digunakan terbatas pada data publik yang tersedia setelah proses *web scraping* dan sesuai dengan kebijakan penggunaan data tokopedia, sehingga tidak termasuk data ulasan yang dihapus atau dibatasi aksesnya oleh platform.
3. Model akan mengklasifikasikan sentimen ulasan berdasarkan tiga kategori, yaitu: Positif, Netral, dan Negatif
4. Hanya menggunakan ulasan dalam Bahasa Indonesia, tanpa menerjemahkan atau menganalisis ulasan berbahasa asing.
5. Produk yang diteliti terbatas pada kategori serum wajah dari sepuluh *brand skincare* lokal, yaitu Somethinc, Skintific, Avoskin, Glad2Glow, Scarlett, The Originote, Azarine, NPURE, Whitelab, dan Wardah, dengan 500 ulasan per *brand*.
6. Pada penelitian ini, model yang digunakan terbatas pada model IndoBERT (indobenchmark/indobert-base-p1) tanpa membandingkan performanya

dengan model Transformer lain seperti mBERT atau XLM-R.

7. Teknik augmentasi data yang diterapkan terbatas pada *back-translation* (Bahasa Indonesia $\rightarrow$ Bahasa Inggris $\rightarrow$ Bahasa Indonesia) dan *synonym replacement*, dengan asumsi bahwa kedua teknik ini cukup efektif untuk memperkaya variasi *dataset training*.
8. Jumlah *dataset* yang digunakan adalah 5.000 ulasan asli (500 ulasan per *brand* $\times$ 10 *brand*) sebelum proses augmentasi, dan data akan dibagi menjadi 70% *training*, 15% *validation*, dan 15% *test set*.
9. Kinerja model akan dievaluasi dengan memakai metrik seperti: *Accuracy*, *Precision*, *Recall*, dan *F1-score*.