BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Sebelum melakukan penelitian mengenai sistem rekomendasi produk kecantikan lokal berbasis konten dan menggunakan *Natural Language Processing*, penulis menggunakan beberapa studi penelitian sebelumnya yang berhubungan sebagai referensi dan pedoman untuk penelitian ini. Berikut beberapa studi penelitian sebelumnya:

Al Kautsar dan Hasrullah [3] dalam penelitiannya yang mengembangkan sistem rekomendasi berbasis konten (*Content-Based Filtering*) untuk pemilihan produk *skincare* dengan memanfaatkan data ulasan pengguna dan komposisi bahan (*ingredients*) yang ada di *website* Sociolla. Penelitian tersebut menerapkan algoritma TF-IDF untuk pembobotan kata dan *Cosine Similarity* untuk mengukur tingkat kemiripan antar produk, di mana hasil analisis menunjukkan bahwa sistem mampu memberikan rekomendasi yang memiliki relevansi tinggi dengan preferensi pengguna serta efektif mengurangi risiko kesalahan pemilihan produk. Sejalan dengan pendekatan tersebut, penelitian Idris dan Pontoiyo [16] juga menerapkan metode *Content-Based Filtering* dengan kombinasi TF-IDF dan *Cosine Similarity*, berfokus pada rekomendasi produk *makeup* untuk mencari alternatif produk ketika produk utama tidak tersedia. Evaluasi kinerja sistem ini menggunakan metrik *Mean Average Precision* (MAP) dan menghasilkan nilai sempurna (1,0), yang mengindikasikan bahwa produk yang direkomendasikan sangat relevan berdasarkan kemiripan kategori dan deskripsi produk.

Selain pendekatan berbasis konten teks, sistem rekomendasi produk kecantikan juga banyak dikembangkan menggunakan pendekatan *Multi-Criteria Decision Making* (MCDM). Salah satunya adalah penelitian oleh Sulistiyorini [15], yang menerapkan metode *Fuzzy Analytical Hierarchy Process* (FAHP) pada sistem pendukung keputusan pemilihan *sunscreen* untuk kulit kering. Penelitian tersebut menggunakan empat kriteria statis, yaitu SPF, harga, netto, dan kemasan. Penggunaan metode FAHP dengan representasi skala *Triangular Fuzzy Number* (TFN) dalam studi ini terbukti mampu menutupi kelemahan AHP konvensional dengan mengurangi tingkat subjektivitas

penilaian kriteria. Sistem tersebut berhasil mencapai tingkat akurasi sebesar 83% yang termasuk dalam kategori 'Sangat Baik'. Meskipun demikian, pendekatan MCDM seperti ini memiliki batasan karena sangat bergantung pada kriteria terstruktur yang ditentukan secara manual dan belum mampu memproses informasi tidak terstruktur seperti konteks deskripsi produk secara otomatis.

Di luar produk kecantikan, terdapat pula sistem rekomendasi resep makanan berdasarkan bahan makanan yang tersedia berbasis aplikasi *mobile* oleh Andina, Pramukantoro, dan Yahya [19]. Dalam penelitian ini, dibandingkan performa metode representasi teks, yaitu TF-IDF dan IndoBERT, dalam hal relevansi hasil rekomendasi dan efisiensi waktu respon. Metrik evaluasi yang digunakan yaitu *Precision@5, Recall@5, F1-score@5,* dan *Mean Reciprocal Rank* (MRR@5) untuk mengukur relevansi hasil. Hasil penelitian menunjukkan bahwa IndoBERT unggul secara relevansi, dengan skor rata-rata yang lebih tinggi pada semua metrik evaluasi dibandingkan TF-IDF. Namun, TF-IDF memiliki keunggulan dari sisi dalam waktu respon yang lebih cepat. Selain itu, terdapat juga penelitian oleh Silfianti *et al*. [17] yang mengembangkan sistem rekomendasi destinasi wisata dengan memanfaatkan penggabungan IndoBERT dan TF-IDF di tingkat skor *similarity* akhir, serta *fine-tuning* dengan SimCSE (*Simple Contrastive Sentence Embedding*). Pendekatan *hybrid* ini berhasil menangkap makna semantik melalui IndoBERT sekaligus mempertahankan relevansi kata kunci literal melalui TF-IDF. Hasil evaluasi menunjukkan performa sistem yang optimal, dengan nilai *Precision, Recall, F1-Score* mencapai 1,0 pada *threshold* tertentu.

Pada ranah global pun, pendekatan *hybrid* berbasis *dense vector* dan *sparse vector* di tingkat fitur/vektor telah diterapkan pada sistem rekomendasi produk terkait *K-wave* (*Korean Wave*), seperti pakaian, kosmetik, hingga elektronik Korea [18]. Penelitian oleh Lee *et al*. tersebut menggabungkan representasi *dense vector* berbasis BERT dengan *sparse vector* berbasis TF-IDF untuk memahami hubungan konteks antarproduk sekaligus mempertahankan kecocokan kata kunci. Hasil penelitian menunjukkan bahwa pendekatan *ensemble* menghasilkan performa yang lebih baik dibandingkan penggunaan dense vector maupun *sparse vector* secara individual. Pada kategori elektronik, kombinasi bobot 0,3 untuk BERT dan 0,7 untuk TF-IDF menghasilkan performa terbaik dengan nilai *Mean Reciprocal Rank* (MRR) tertinggi sebesar 0,56825.

Penelitian tersebut juga menyimpulkan bahwa penggabungan *dense vector* dan *sparse vector* mampu meningkatkan kualitas rekomendasi karena dapat memanfaatkan informasi semantik dan relevansi tekstual secara bersamaan.

Berdasarkan beberapa penelitian terdahulu tersebut, sebagian besar sistem rekomendasi produk kecantikan masih menggunakan metode berbasis *Multi-Criteria Decision Making* (MCDM) dengan pembobotan statis atau teknik ekstraksi fitur yang relatif sederhana. Meskipun beberapa penelitian telah menggunakan pendekatan *hybrid* TF-IDF dan IndoBERT untuk meningkatkan relevansi rekomendasi, penerapan metode tersebut pada domain produk kecantikan lokal serta integrasi aspek legalitas produk, seperti BPOM, masih terbatas. Oleh karena itu, penelitian ini mengusulkan pengembangan sistem rekomendasi yang mampu memanfaatkan informasi produk secara lebih kontekstual dan mempertimbangkan aspek legalitasnya untuk menghasilkan rekomendasi produk kecantikan lokal yang lebih relevan.

Berdasarkan beberapa penelitian terdahulu tersebut, sistem rekomendasi produk kecantikan yang ada umumnya masih terbatas pada satu merek atau *platform* tertentu [3], maupun satu jenis produk saja [18], sehingga belum merepresentasikan keragaman produk kecantikan lokal secara menyeluruh. Dari sisi pendekatan, sebagian besar masih mengandalkan metode berbasis statistik seperti TF-IDF dan *Cosine Similarity* yang terbatas dalam menangkap makna kontekstual dan nuansa bahasa [3][12]. Meskipun beberapa penelitian telah mengeksplorasi pendekatan berbasis *transformer* seperti IndoBERT [14][15] maupun *hybrid* BERT dan TF-IDF [13] dengan hasil yang lebih baik secara semantik, penerapannya pada domain produk kecantikan lokal Indonesia masih sangat terbatas. Selain itu, belum ada yang mengintegrasikan aspek keamanan dan legalitas produk, seperti status pendaftaran BPOM pada sistem. Oleh karena itu, penelitian ini mengusulkan pengembangan sistem rekomendasi produk kecantikan lokal yang memanfaatkan pendekatan berbasis konten dengan NLP untuk memahami informasi produk secara lebih kontekstual, sekaligus mengintegrasikan aspek legalitas BPOM.

Tabel 2. 1 Penelitian Terdahulu

No.	Peneliti	Judul Penelitian	Tahun	Tujuan	Metode	Hasil
1.	Al Kautsar & Hasrullah	Sistem Rekomendasi Pemilihan Produk *Skincare*	2025	Mengembangkan sistem rekomendasi produk *skincare* yang tepat berdasarkan ulasan pengguna dan komposisi bahan (*ingredients*).	*Content-Based Filtering*, TF-IDF, *Cosine Similarity*	Sistem mampu memberikan rekomendasi produk yang memiliki nilai kemiripan tinggi (skor *Cosine Similarity* tertinggi 0,411) dengan preferensi pengguna
2.	Idris & Pontoiyo	Sistem Rekomendasi Produk *Makeup* Berbasis *Content-Based Filtering* dengan TF-IDF dan *Cosine Similarity*	2025	Membangun sistem untuk merekomendasikan produk alternatif yang relevan ketika produk *makeup* utama tidak tersedia.	*Content-Based Filtering*, TF-IDF, *Cosine Similarity*	Sistem berhasil merekomendasikan 5 produk alternatif yang relevan dengan nilai evaluasi *Mean Average Precision* (MAP) sempurna yaitu 1,0

No.	Peneliti	Judul Penelitian	Tahun	Tujuan	Metode	Hasil
3.	Sulistyorini	Penerapan Metode *Fuzzy* AHP pada Sistem Rekomendasi Pemilihan *Sunscreen*	2025	Menentukan rekomendasi pemilihan *sunscreen* untuk kulit kering secara akurat berdasarkan 4 kriteria pengambil keputusan (SPF, harga, netto, kemasan).	*Multi-Criteria Decision Making* (MCDM) dengan *Fuzzy Analytical Hierarchy Process* (FAHP).	Sistem mampu menutupi kelemahan subjektivitas AHP konvensional menggunakan skala TFN dan memberikan hasil rekomendasi dengan tingkat akurasi 83% (kategori "Sangat Baik").
4.	Andina, Pramukan-toro, & Yahya	Studi Kinerja TF-IDF dan IndoBERT dalam Rekomendasi Resep Berbasis *Mobile* dengan Arsitektur *Microservice*	2025	Mengevaluasi sistem rekomendasi resep berbasis *mobile* dengan membandingkan TF-IDF dan IndoBERT dalam hal relevansi hasil rekomendasi dan efisiensi waktu respon.	*Cosine Similarity,* TF-IDF, IndoBERT, Arsitektur MVVM (Model-View-ViewModel)	IndoBERT unggul dalam hal relevansi, dengan skor rata-rata yang lebih tinggi pada semua metrik evaluasi dibandingkan TF-IDF. Namun, TF-IDF memberikan waktu respon yang secara signifikan lebih cepat dibandingkan IndoBERT.

No.	Peneliti	Judul Penelitian	Tahun	Tujuan	Metode	Hasil
5.	Silfianti *et al*.	*A Deep Learning Approach for Tourism Destination Recommendation Using IndoBERT and TF-IDF*	2025	Mengusulkan sistem rekomendasi destinasi pariwisata di Indonesia yang terpersonalisasi dengan menganalisis deskripsi destinasi secara mendalam.	CBF *Hybrid* yang menggabungkan model IndoBERT *SimCSE,* TF-IDF, *Cosine Similarity*	Komposisi IndoBERT dan TF-IDF memberikan keseimbangan optimal antara pencocokan semantik dan leksikal, mencapai nilai 1,0 untuk *Precision, Recall,* dan *F1-score* pada ambang batas (*threshold*) kemiripan 0,20.
6.	Lee *et al*.	*Recommending K-Wave Items Tailored for Small-Sized Exporters by Incorporating Dense and Sparse Vectors*	2023	Membangun sistem pendukung keputusan untuk eksportir skala kecil guna menemukan tren item *K-Wave* (seperti kosmetik dan elektronik) di negara target.	*Ensemble* CBF menggabung-kan *KOBERT (dense vector)* dan TF-IDF *(sparse vector), Cosine Similarity*	Metode *hybrid* (*ensemble*) mengungguli metode dasar tunggal. Khusus untuk dataset kategori kosmetik, akurasi tertinggi dicapai menggunakan kombinasi bobot 0,3 untuk BERT dan 0,7 untuk TF-IDF.

Tabel 2. 2 Perbandingan Metode

No.	Peneliti	Pendekatan	Metode	Kelebihan	Kekurangan
1.	Al Kautsar & Hasrullah	*Content-Based Filtering*	TF-IDF & *Cosine Similarity*	Sederhana, komputasi ringan, dan efektif mencocokkan kemiripan ulasan/deskripsi produk tanpa bergantung pada data interaksi pengguna (*cold start*).	TF-IDF bersifat statis, hanya menghitung frekuensi kata tanpa memahami konteks kalimat. Jika dua produk menggunakan sinonim, TF-IDF akan menganggapnya berbeda.
2.	Idris & Pontoiyo	*Content-Based Filtering*	TF-IDF & *Cosine Similarity*	Sederhana, komputasi ringan, mampu memberikan rekomendasi produk makeup pengganti lintas merek dengan tingkat evaluasi relevansi (MAP) sangat tinggi secara otomatis.	TF-IDF masih terbatas pada pencocokan kata persis. Tidak mampu memahami nuansa bahasa kompleks atau bahasa lokal konsumen. Sinonim bahasa dianggap berbeda.

No.	Peneliti	Pendekatan	Metode	Kelebihan	Kekurangan
3.	Sulistyorini	*Content-Based Filtering* (MCDM)	*Fuzzy Analytical Hierarchy Process* (FAHP)	Mampu mengurangi tingkat subjektivitas penilaian kriteria konvensional menggunakan skala *Triangular Fuzzy Number* (TFN).	Sangat bergantung pada kriteria statis yang ditentukan manual dan tidak mampu memproses informasi tidak terstruktur seperti teks deskripsi produk secara otomatis
4.	Andina, Pramukan-toro, & Yahya	*Content-Based Filtering*	Komparasi TF-IDF vs IndoBERT & *Cosine Similarity*	Membuktikan secara empiris bahwa IndoBERT mengungguli TF-IDF dalam segala metrik relevansi (*Precision, Recall, F1-Score, MRR*) karena pemahaman kontekstualnya	IndoBERT memerlukan waktu komputasi lebih tinggi (367ms) dibanding TF-IDF (105 ms), karena komputasi model *transformer* yang kompleks.

No.	Peneliti	Pendekatan	Metode	Kelebihan	Kekurangan
5.	Silfianti *et al.*	*Hybrid Content-Based Filtering*	*Weighted Composition* IndoBERT + SimCSE + TF-IDF menggunakan *Cosine Similarity*	IndoBERT sangat akurat dalam memahami konteks semantik dan linguistik bahasa Indonesia. Penambahan SimCSE secara signifikan meningkatkan kualitas representasi di level kalimat, dan TF-IDF mengekstrak kata kunci esensial, menciptakan penggabungan yang sangat seimbang.	Sangat sensitif terhadap nilai ambang batas (*threshold*) *cosine similarity*. Penggunaan *threshold* yang lebih ketat/tinggi (misalnya di atas 0,25 atau 0,27) menyebabkan nilai *Recall* dan *F1-Score* turun secara drastis.
6.	Lee *et al*.	*Ensemble Content-Based Filtering*	*Weighted Feature-Level* KOBERT (*Bidirectional Encoder Representations from Transformers*) + TF-IDF menggunakan *Cosine Similarity*	Mampu menangkap informasi relasi kontekstual antar item yang tidak terdeteksi klasifikasi konvensional sekaligus mempertahankan keakuratan pencocokan kata kunci literal.	Masih terbatas pada kategori tertentu (pakaian, kosmetik, elektronik) dan hanya memodelkan konteks berdasarkan format teks/berita, tanpa mempertimbangkan sumber informasi ekosistem lain seperti data media sosial.

2.2. Dasar Teori

2.2.1. Produk Kecantikan

Produk kecantikan atau kosmetik merupakan zat atau bahan yang digunakan pada bagian luar tubuh (seperti kulit, rambut, dan kuku) dengan tujuan untuk membersihkan, mempercantik, meningkatkan daya tarik, atau mengubah penampilan. Saat ini, pasar produk kecantikan di Indonesia mengalami perkembangan yang sangat pesat dan sangat digemari oleh masyarakat karena produknya dianggap memiliki kandungan yang bermanfaat bagi kesehatan kulit, seperti kemampuan untuk mencerahkan wajah dalam waktu yang relatif cepat [20]. Industri lokal pun turut serta mengambil bagian dalam pasar produk kecantikan ini, terlebih dengan menginovasikan formula dan penyesuaian pada karakteristik jenis kulit, warna kulit masyarakat Indonesia, juga kondisi iklim.

Namun, popularitas ini juga diikuti dengan munculnya tantangan serius bagi konsumen. Terdapat banyak pelaku usaha yang memanfaatkan tingginya permintaan pasar terutama akan produk yang memiliki hasil instan, yakni dengan menjual produk kosmetik yang mengandung bahan berbahaya atau tidak memiliki izin edar resmi dari Badan Pengawas Obat dan Makanan (BPOM) demi mendapatkan keuntungan besar [13]. Peredaran kosmetik ilegal ini sangat berisiko menimbulkan kerugian bagi penggunanya, yakni menyangkut kualitas, mutu barang, informasi yang tidak jelas bahkan menyesatkan, dan pemalsuan [21].

Selain aspek legalitas, keamanan penggunaan produk kecantikan tidak juga sangat dipengaruhi oleh kecocokan dengan jenis kulit masing-masing individu. Setiap orang memiliki karakteristik kulit wajah yang berbeda-beda, seperti kulit normal, kering, berminyak, atau sensitif. Sehingga, kandungan bahan dalam produk kecantikan, termasuk produk lokal perlu disesuaikan dengan kebutuhan kulit pengguna. Ketidaksesuaian antara komposisi bahan dan kondisi kulit dapat memicu berbagai reaksi negatif, seperti iritasi, kemerahan, hingga alergi yang serius [13]. Oleh karena itu, pemahaman mengenai komposisi bahan,

jenis kulit, dan status keamanan produk menjadi faktor krusial dalam memitigasi risiko kesehatan bagi konsumen produk kecantikan.

2.2.2. Badan Pengawas Obat dan Makanan (BPOM)

Badan Pengawas Obat dan Makanan (BPOM) merupakan lembaga pemerintah non-kementerian yang bertanggung jawab atas pengawasan obat dan makanan di Indonesia, termasuk produk kosmetik. BPOM menjalankan fungsi pengawasan berdasarkan berbagai regulasi yang mengatur tentang kosmetik, yang terus berubah dan berkembang. Salah satu regulasi terbaru adalah Peraturan BPOM Nomor 16 Tahun 2024 tentang Batas Cemaran dalam Kosmetik untuk menyelaraskan standar nasional dengan standar internasional guna meningkatkan keamanan produk dan melindungi konsumen dari risiko kesehatan yang terkait dengan bahan berbahaya dalam kosmetik [8].

Berdasarkan Keputusan Presiden Nomor 103 Tahun 2001 Pasal 67 C, BPOM harus berkoordinasi dengan semua pihak dalam mengawasi peredaran kosmetik ilegal. Ketentuan ini menunjukkan bahwa pengawasan tidak dapat dibebankan kepada BPOM semata, melainkan membutuhkan partisipasi aktif pelaku usaha, *platform e-commerce*, dan konsumen [22].

2.2.3. *E-Commerce*

Electronic Commerce atau *e-commerce* merupakan model perniagaan yang yang memungkinkan perusahaan maupun individu melakukan aktivitas jual beli barang dan jasa melalui internet. Data menunjukkan bahwa sekitar 67% pengguna di Indonesia memilih aplikasi *e-commerce* sebagai saluran utama dalam berbelanja, menunjukkan tingginya adopsi teknologi digital dalam kehidupan sehari-hari [1]. Perubahan ini tidak hanya berdampak pada mekanisme transaksi, tetapi juga memengaruhi pemasaran, dinamika pasar, serta pola interaksi antara produsen dan konsumen [23].

Shopee menjadi salah satu platform e-commerce yang menonjol di Indonesia melalui penerapan strategi pemasaran digital inovatif, khususnya integrasi fitur gamifikasi seperti permainan interaktif, sistem poin, e-wallet, voucher belanja, dan layanan pinjaman digital. Strategi ini terbukti mampu

meningkatkan tingkat keterlibatan (*engagement*) dan kepuasan pelanggan secara signifikan, sehingga berkontribusi pada peningkatan loyalitas [24].

Dalam sektor produk kecantikan, pemasaran melalui media sosial melalui ulasan *influencer*, *affiliate*, atau pengguna secara *online* yang disajikan dalam format visual dan narasi personal, atau yang disebut *User Generated Content* berperan besar dalam membentuk persepsi dan keputusan pembelian konsumen. Ulasan ini dianggap konsumen lebih autentik dibandingkan iklan konvensional [25]. Ketergantungan konsumen terhadap ulasan *influencer* yang mencapai 80,7% sebagai sumber informasi utama menyebabkan perhatian terhadap status izin edar BPOM dan kandungan produk menjadi terabaikan [26].

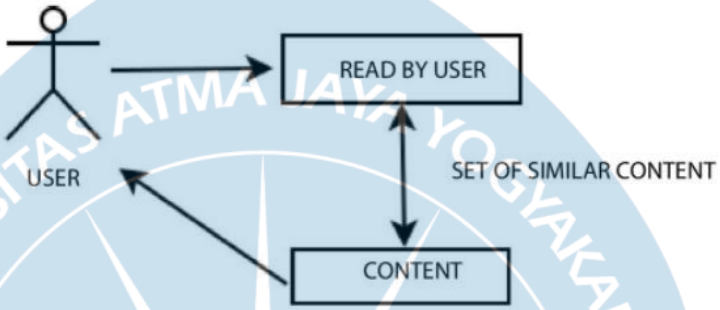
Kondisi ini diperparah dengan maraknya produk kosmetik impor yang viral di luar negeri namun belum memiliki izin edar BPOM, sehingga kecepatan penyebaran tren kerap melampaui proses verifikasi resmi dan memungkinkan produk bermasalah tetap beredar luas di platform *e-commerce* [27].

2.2.4. Sistem Rekomendasi

Sistem rekomendasi adalah teknologi yang dirancang untuk memberikan saran atau rekomendasi objek kepada pengguna berdasarkan algoritma tertentu, dengan tujuan utama meningkatkan pengalaman pengguna dan mempermudah pengambilan keputusan. Sistem rekomendasi pertama kali diperkenalkan dari studi Belkin dan Croft pada tahun 1992, yang membedakan penyaringan informasi (*information filtering*) dan pengambilan informasi (*information retrieval*). Studi tersebut kemudian memicu lahirnya sistem "Tapestry" pada tahun yang sama sebagai sistem penyaringan kolaboratif (*Collaborative Filtering*) pertama [28]. Kini, sistem rekomendasi berkembang pesat melalui tiga pendekatan utama yakni *Collaborative Filtering*, *Content-Based Filtering*, dan *Hybrid*.

Dalam merekomendasikan suatu item, sistem rekomendasi bekerja dengan menganalisis pola kebiasaan pengguna serta memodelkan data untuk menyarankan item yang paling relevan, sehingga menjadi solusi efektif untuk mengatasi masalah kelebihan informasi (*information overload*) di tengah

banyaknya pilihan produk yang tersedia di pasar. Lebih spesifik, pendekatan *Content-Based Filtering* memberikan saran item yang relevan berdasarkan kemiripan karakteristik atau atribut item tersebut dengan preferensi masa lalu pengguna tanpa bergantung pada data personal pengguna lain. Pendekatan ini banyak digunakan pada sistem yang memiliki data teks yang kaya, seperti dokumen digital, deskripsi produk, atau katalog. Ilustrasi cara kerja pendekatan *Content-Based Filtering* tersebut seperti pada Gambar 2.1.



Gambar 2. 1 Ilustrasi *Content-Based Filtering*

2.2.5. Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah bidang ilmu yang menggabungkan ilmu komputer, kecerdasan buatan, dan psikologi kognitif. NLP adalah teknologi yang memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa manusia secara bermakna. Berbeda dengan data yang terstruktur, NLP menangani bahasa alami yang tidak teratur dan sering kali memiliki banyak makna, sehingga diperlukan teknik khusus untuk memahami maksud dari teks atau ucapan.

NLP menggunakan berbagai teknik dari kecerdasan buatan seperti pembelajaran mesin (*machine learning*) untuk menjalankan tugas-tugas yang melibatkan pemahaman bahasa. Sejak tahun 2000-an, pembelajaran mesin menjadi sangat populer dalam NLP karena dapat membangun sistem yang cerdas dengan lebih efisien dibandingkan sistem manual yang membutuhkan banyak aturan dari para ahli [29]. Dalam perkembangannya, terdapat beberapa pendekatan utama dalam pemrosesan teks menggunakan NLP, yaitu berbasis statistik seperti TF-IDF, berbasis *word embedding* seperti Word2Vec dan FastText,

hingga berbasis *deep learning* seperti model Transformer yang menjadi dasar dari BERT (*Bidirectional Encoder Representations from Transformers*).

Pada pengolahan teks menggunakan model NLP, tahapan *text preprocessing* menjadi langkah paling penting untuk memastikan kualitas data dan representasi teks. Tahapan tersebut seperti *case folding* (mengubah huruf besar ke kecil), *tokenizing* (memecah kalimat menjadi kata), *stopword removal/filtering* (menghapus kata umum yang tidak bermakna), dan *stemming* (mengubah kata ke bentuk dasar). Proses ini bertujuan untuk mengurangi *noise* pada teks dan meningkatkan efektivitas analisis oleh model nantinya.

2.2.6. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode representasi teks yang memberikan bobot pada sebuah kata berdasarkan frekuensinya di dalam suatu dokumen dan tingkat kelangkaannya di seluruh kumpulan dokumen [16]. Dalam konteks sistem rekomendasi, TF-IDF digunakan untuk mengevaluasi seberapa penting suatu istilah (*term*) yang terdapat dalam karakteristik item atau profil pengguna, relatif terhadap seluruh katalog item atau kumpulan pengguna di dalam sistem.

Pada metode TF-IDF, pembobotan yang lebih tinggi diberikan pada istilah-istilah yang paling jarang muncul secara keseluruhan, karena istilah tersebut dianggap sebagai faktor pembeda (*discriminating factors*) yang paling kuat. Bobot akhir untuk sebuah istilah diperoleh dengan mengalikan kedua ukuran tersebut, yang diformulasikan sebagai berikut:

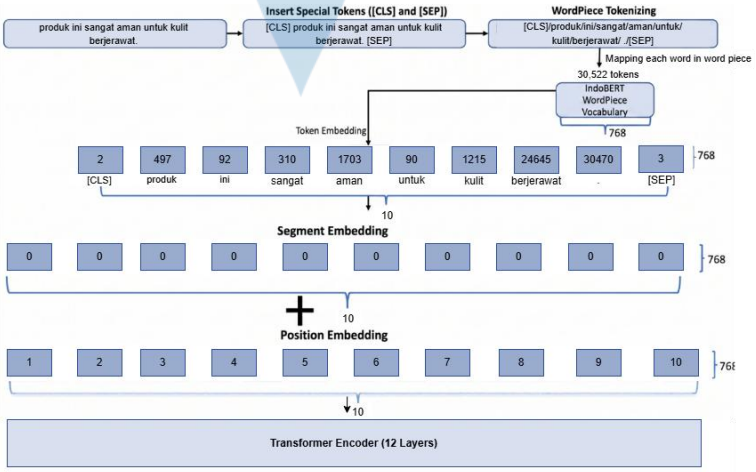
$$tfidf_{k,j} = tf_{k,j} \times idf_k$$

Vektor-vektor baru yang merepresentasikan item atau pengguna dari hasil perhitungan TF-IDF ini kemudian dapat digunakan sebagai dasar untuk perhitungan ukuran kemiripan (*similarity measure*) yang melibatkan perbandingan vektor, seperti algoritma *Cosine Similarity* [30].

2.2.7. *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT)

IndoBERT adalah model bahasa *pre-trained* (telah dilatih sebelumnya) yang dikembangkan khusus untuk pemrosesan bahasa alami (NLP) dalam bahasa Indonesia. Model ini dibangun berdasarkan arsitektur BERT (*Bidirectional Encoder Representations from Transformers*) yang diperkenalkan oleh Google. IndoBERT dikembangkan oleh Koto, Rahimi, Lau, dan Baldwin pada tahun 2020 sebagai respons atas minimnya sumber daya dan dataset terstandarisasi untuk riset NLP bahasa Indonesia, meskipun bahasa ini memiliki penutur yang sangat besar yakni sekitar 200 juta jiwa. Sebelum adanya IndoBERT, penelitian NLP bahasa Indonesia sering kali bergantung pada model multilingual (seperti mBERT) atau model monolingual lain yang belum sepenuhnya menangkap nuansa bahasa Indonesia secara optimal.

Cara kerja IndoBERT berfokus pada mekanisme *Masked Language Modeling* (MLM). Berbeda dengan model sekuensial tradisional yang membaca teks satu arah (kiri ke kanan) contohnya seperti LSTM, IndoBERT ini menggunakan mekanisme *transformer* yang memungkinkan model untuk menelaah seluruh kata dalam kalimat secara bersamaan dari dua arah (*bidirectional*). Dalam proses pelatihannya, sebagian kata dalam kalimat masukan ditutupi (*masked*), dan model diuji untuk memprediksi kata yang hilang tersebut berdasarkan konteks kata-kata di sekitarnya. Arsitektur cara kerjamodel IndoBERT seperti pada Gambar 2.2.



Gambar 2. 2 Arsitektur Representasi IndoBERT

Pendekatan ini memungkinkan IndoBERT untuk menghasilkan representasi vektor (*embedding*) yang kaya akan makna semantik dan sintaksis, serta terbukti mencapai performa *state-of-the-art* pada berbagai tugas NLP bahasa Indonesia yang telah diberikan untuk diuji, seperti penandaan *Part-of-Speech* (POS), pengenalan entitas bernama (NER), dan analisis sentimen, menunjukkan IndoBERT bahkan lebih unggul daripada model multilingual BERT (mBERT) maupun model berbasis *embedding* statis [31].

2.2.8. *Cosine Similarity*

Cosine Similarity merupakan metrik untuk mengukur kemiripan dokumen yang paling umum. Metrik ini mengukur kemiripan antara dua vektor *non-zero* dalam ruang *inner product* dengan menghitung kosinus dari sudut di antara keduanya. Dalam konteks sistem rekomendasi, *item* dan *user* direpresentasikan dalam bentuk vektor. Representasi vektor ini menyederhanakan perhitungan dengan membatasi diri pada karakteristik yang dapat dihitung dan mudah diperhitungkan, menunjukkan *keyword* dalam *profile item* atau *user*. Nilai 1 akan digunakan untuk menunjukkan keberadaan *keyword* dalam karakteristik *item* atau profil *user*, dan nilai 0 akan digunakan untuk menunjukkan ketidakhadirannya. Secara matematis, *Cosine Similarity* dihitung menggunakan rumus:

$$\textit{cosine similarity} = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{|\,\boldsymbol{A}\,||\,\boldsymbol{B}\,|}$$

Di mana $\mathbf{A} \cdot \mathbf{B}$ menunjukkan *dot product* dari vektor $\mathbf{A}$ dan $\mathbf{B}$, sedangkan $|\,\mathbf{A}\,|$ dan $|\,\mathbf{B}\,|$ merepresentasikan *Euclidean norm* mereka. *Dot product* dihitung dengan mengalikan elemen-elemen yang bersesuaian dari vektor dan menjumlahkan hasilnya, sementara *Euclidean norm* dari sebuah vektor adalah akar kuadrat dari jumlah kuadrat elemen-elemennya. Dalam kasus dua item $i_1$dan i_2, rumusnya adalah sebagai berikut:

$$sim_{cosine}(i_1, i_2) = \frac{\overrightarrow{i_1} \cdot \overrightarrow{i_2}}{||\,\overrightarrow{i_1}\,|| \cdot ||\,\overrightarrow{i_2}\,||}$$

Nilai *Cosine Similarity* yang dihasilkan biasanya dinormalisasi untuk berada dalam rentang -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan

kemiripan yang lebih besar antara vektor. *Cosine Similarity* sangat efektif untuk menangani vektor sparse dan berdimensi tinggi, yaitu seperti vektor *term-frequency* dalam analisis teks, karena fokus pada elemen *non-zero* dan mengabaikan *zero-matches* yang umum dalam struktur data tersebut [32].

2.2.9. *Weighted Composition* (Pembobotan Komposisi)

Penggabungan kemiripan berbobot (*weighted similarity*) pada tingkat skor merupakan teknik dalam sistem rekomendasi *hybrid* yang mengkalkulasi nilai kemiripan dari masing-masing model secara terpisah sebelum digabungkan menjadi satu kesatuan [17]. Setelah kedua nilai kemiripan (*similarity score*) tersebut didapatkan dari masing-masing metode, hasilnya diintegrasikan melalui komposisi berbobot (*weighted composition*). Teknik ini bertujuan untuk menggabungkan dan menyeimbangkan kelebihan dari berbagai karakteristik representasi data, sehingga menghasilkan sistem rekomendasi yang lebih akurat.

Secara matematis, penggabungan di tingkat skor akhir ini diimplementasikan menggunakan kombinasi linier berbobot (*weighted linear combination*). Formula untuk menghitung skor akhir dapat dituliskan sebagai berikut:

$$Score = w_1 \cdot sim_A + w_2 \cdot sim_B$$

di mana w_1 dan w_2 merupakan nilai bobot yang telah ditentukan untuk menyelaraskan persentase pengaruh dari masing-masing *similarity* model. Pendekatan perhitungan pada tahap skor akhir ini memberikan fleksibilitas untuk menyesuaikan parameter bobot dengan karakteristik dataset yang digunakan, sehingga sistem dapat menyeimbangkan berbagai aspek penilaian untuk menghasilkan rekomendasi yang lebih menyeluruh dan relevan [33].

2.2.10. Evaluasi Model

Tahap evaluasi merupakan tahap untuk melakukan penilaian performa model sistem rekomendasi yang telah dibuat, yang dilakukan dengan mengukur kemampuan model dalam menghasilkan rekomendasi relevan menggunakan dataset uji. Evaluasi ini memberikan gambaran objektif tentang kemampuan model pada situasi dunia nyata dan mengetahui sejauh mana model dapat

diandalkan dalam tugas yang diberikan [4]. Pada tahap ini, akan diterapkan metrik evaluasi untuk menilai kualitas model atau algoritma pembelajaran mesin (*machine learning*) yaitu *precision*@K, *F1-score*@K, MAP@K, dan NDCG@K. Masing-masing metrik ini memiliki rumusan khusus, yaitu:

a. ***Precision*@K**

$$Precision@\mathrm{K} = \frac{True\ Positive\ (\mathrm{TP})}{K}$$

b. ***F1-score*@K**

$$F1 - score@\mathrm{K} = 2 \times \frac{Precision@\mathrm{K} \times Recall@K}{Precision@\mathrm{K} + Recall@K}$$

c. **MAP@K (*Mean Average Precision at K*)**

MAP@K dihitung dalam dua tahap:

$$\mathrm{AP@K} = \frac{1}{\min(K, R)} \sum_{k=1}^{K} P(k) \cdot \mathrm{rel}(k)$$

$$\mathrm{MAP@K} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \mathrm{AP@K}(q)$$

d. **NDCG@K (*Normalized Discounted Cumulative Gain at K*)**

NDCG@K dihitung dalam tiga tahap:

$$\mathrm{DCG@K} = \sum_{k=1}^{K} \frac{rel_k}{\log_2(k+1)}$$

$$\mathrm{IDCG@K} = \sum_{k=1}^{K} \frac{rel_k^{\mathrm{ideal}}}{\log_2(k+1)}$$

$$\mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}$$

Di mana P(k) adalah *Precision* pada posisi ke-*k*, rel(*k*) bernilai 1 jika item pada posisi ke-*k* relevan dan 0 jika tidak, *R* adalah total item relevan, dan *Q* adalah himpunan seluruh *query*/pengguna yang dievaluasi. Untuk NDCG@K, rel_i adalah skor relevansi item pada posisi ke-*i*, $\log_2(i+1)$ adalah faktor diskon logaritmik, dan IDCG@K adalah nilai DCG dari urutan rekomendasi yang sempurna. *K* pada

seluruh metrik merujuk pada jumlah total item teratas yang diprediksi sistem sebagai relevan atau positif [33].

Selain rumusan matematis di atas, perhitungan metrik evaluasi sistem rekomendasi sangat bergantung pada *ground truth*. *Ground truth* merupakan data aktual yang dibutuhkan sebagai acuan utama untuk mengevaluasi sejauh mana hasil rekomendasi dari model sesuai dengan preferensi nyata pengguna. Tanpa adanya data *ground truth*, sistem tidak memiliki tolak ukur yang valid untuk menilai apakah item yang direkomendasikan benar-benar relevan atau hanya bersifat acak [34].

2.2.11. Streamlit

Streamlit adalah *framework open source* populer terutama di kalangan pengembang *data science* dan *machine learning* untuk membuat antarmuka *website* atau aplikasi *website* yang responsif hanya dengan menggunakan Python. Streamlit memudahkan perancangan antarmuka dengan menyediakan visualisasi data, tabel, grafik, dan lain sebagainya. Secara keseluruhan, kemampuan *framework* ini dapat menyederhanakan proses *deployment website,* sehingga memudahkan untuk mengaksesnya secara *online* [35].