BAB II

TINJAUAN PUSTAKA

2.1 Penelitian yang Relevan

Sebagai bahan perbandingan dengan penelitian yang akan dilakukan, bagian ini akan mencantumkan beberapa penelitian yang telah dilakukan oleh peneliti sebelumnya dengan topik dan penggunaan metode yang sama dan hampir mirip serta pengembangan teori yang berkaitan pada penelitian ini.

2.1.1 Matriks Penelitian

Tabel 2. 1 Matriks Penelitian

No	Nama Penelitian	Tri Buwono Bagus Wicaksono & Rama Dian Syah
1	Judul Jurnal	Implementasi Metode Bidirectional Encoder Representations from Transformers untuk Analisis Sentimen terhadap Ulasan Aplikasi Access
	Jenis Data	Data ulasan pengguna aplikasi Access by KAI berbahasa Indonesia
	Sumber Data	Dataset ulasan pengguna dari Google Play Store sebanyak 10.000 data, setelah proses data cleaning menjadi 9.260 ulasan
	Metode	Metode Bidirectional Encoder Representations from Transformers (BERT) dengan pre-trained model IndoBERT, menggunakan pendekatan CRISP-DM (Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment)
	Hasil	Model BERT mampu memprediksi sentimen pengguna dengan akurasi 85%. Sentimen positif dan negatif terdeteksi dengan baik, namun masih terdapat ketidaktepatan pada sentimen netral. Model juga berhasil di-deploy dalam bentuk prototipe website.
	Model Evaluasi	Menggunakan Confusion Matrix dengan akurasi keseluruhan sebesar 85%, serta evaluasi terhadap prediksi benar dan salah pada tiga kelas sentimen (positif, netral, negatif).
	Tahun	2023
2	Nama Penelitian	Llinet Benavides-Cesar, Miguel-Ángel Manso-Callejo, & Calimanut-Ionut Cira
	Judul Jurnal	Methodology Based on BERT (Bidirectional Encoder Representations from Transformers) to Improve Solar Irradiance Prediction of Deep Learning Models Trained with Time Series of Spatiotemporal Meteorological Information
	Jenis Data	Data sekunder berupa time series irradiance, meteorologi, dan

		informasi spasio-temporal
	Sumber Data	CyL-GHI Dataset (Castile and León, Spanyol) dan Hawaii Oahu Solar Measurement Grid (Amerika Serikat)
	Metode	Metodologi berbasis BERT untuk imputasi data hilang, dilanjutkan dengan model Deep Learning (ST_CNN_v1, ST_CNN_v2, ST_Dilated_CNN) untuk prediksi irradiance
	Hasil	Penggunaan model BERT meningkatkan akurasi prediksi hingga 3% dibanding metode tradisional. Nilai MAE dan RMSE menurun, sementara Forecast Skill (FS) meningkat secara konsisten, terutama di stasiun SO01.
	Model Evaluasi	MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), FS (Forecast Skill), dan R^2 Score
	Tahun	2025
3	Nama Penelitian	Tri Buwono Bagus Wicaksono & Rama Dian Syah
	Judul Jurnal	Implementasi Metode Bidirectional Encoder Representations from Transformers untuk Analisis Sentimen terhadap Ulasan Aplikasi Access
	Jenis Data	Data ulasan pengguna aplikasi Access by KAI berbahasa Indonesia
	Sumber Data	Dataset ulasan pengguna dari Google Play Store sebanyak 10.000 data, setelah proses data cleaning menjadi 9.260 ulasan
	Metode	Metode Bidirectional Encoder Representations from Transformers (BERT) dengan pre-trained model IndoBERT, menggunakan pendekatan CRISP-DM (Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment)
	Hasil	Model BERT mampu memprediksi sentimen pengguna dengan akurasi 85%. Sentimen positif dan negatif terdeteksi dengan baik, namun masih terdapat ketidaktepatan pada sentimen netral. Model juga berhasil di-deploy dalam bentuk prototipe website.
	Model Evaluasi	Menggunakan Confusion Matrix dengan akurasi keseluruhan sebesar 85%, serta evaluasi terhadap prediksi benar dan salah pada tiga kelas sentimen (positif, netral, negatif).
	Tahun	2023
4	Nama Penelitian	Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, & Tatsuya Harada
	Judul Jurnal	Finding and Generating a Missing Part for Story Completion
	Jenis Data	Data sekunder berupa kumpulan cerita pendek lima kalimat (five-sentence stories)
	Sumber Data	ROCStories Corpus (Mostafazadeh et al., 2016) yang berisi 98.161 cerita (78.528 data latih, 9.816 validasi, 9.817 uji)
	Metode	Pendekatan baru bernama Missing Position Prediction (MPP) untuk memprediksi posisi bagian cerita yang hilang, dikombinasikan dengan Story Completion (SC) menggunakan arsitektur Hierarchical Seq2Seq berbasis Sentence-BERT (SBERT) sebagai sentence

		encoder, GRU sebagai context encoder, dan BERT sebagai language model decoder.
	Hasil	Model MPP dengan GRU context encoder mencapai akurasi rata-rata 52.2%, lebih tinggi dibanding metode max-pool context (35.0%). Dalam evaluasi manusia (200 cerita), model menghasilkan cerita yang dinilai setara atau lebih baik dari cerita asli pada 26% kasus. Model paling akurat dalam mendeteksi bagian hilang di awal atau akhir cerita.
	Model Evaluasi	Akurasi klasifikasi lima kelas untuk MPP, serta human evaluation (pairwise comparison) untuk SC.
	Tahun	2020
5	Nama Penelitian	Shivaji Alaparthi & Manit Mishra
	Judul Jurnal	Bidirectional Encoder Representations from Transformers (BERT): A Sentiment Analysis Odyssey
	Jenis Data	Data sekunder (teks ulasan film berlabel)
	Sumber Data	50.000 ulasan film dari Internet Movie Database (IMDB)
	Metode	Eksperimen komparatif menggunakan empat teknik analisis sentimen: 1. Sent WordNet (unsupervised lexicon-based) 2. Logistic Regression (supervised klasik) 3. LSTM (supervised deep learning) 4. BERT (pre-trained deep learning model)
	Hasil	Model BERT menunjukkan kinerja paling unggul dibandingkan tiga model lainnya dalam klasifikasi sentimen teks.
	Model Evaluasi	Akurasi, Presisi, Recall, dan F1-Score
	Tahun	2020
6	Nama Penelitian	Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, & Peng Jiang
	Judul Jurnal	BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer
	Jenis Data	Data sekunder (data interaksi pengguna)
	Sumber Data	Empat dataset benchmark sistem rekomendasi (data perilaku pengguna dari platform e-commerce)
	Metode	Pengembangan model BERT4Rec, yaitu model rekomendasi sekuensial berbasis deep bidirectional self-attention (Transformer). Model dilatih dengan Cloze task objective untuk memprediksi item yang di-mask berdasarkan konteks kiri dan kanan.
	Hasil	BERT4Rec secara konsisten mengungguli berbagai model sekuensial state-of-the-art seperti RNN dan GRU dalam tugas rekomendasi, berkat kemampuan bidirectional yang menangkap konteks dua arah.
	Model Evaluasi	Performa dievaluasi berdasarkan metrik sistem rekomendasi seperti Hit Ratio (HR) dan Normalized Discounted Cumulative Gain (NDCG).
	Tahun	2019

7	Nama Penelitian	Ying Xiong, Shuai Chen, Qingcai Chen, Jun Yan, & Buzhou Tang
	Judul Jurnal	Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study
	Jenis Data	Data sekunder (teks medis dari catatan kesehatan elektronik/EHR)
	Sumber Data	Dataset ClinicalSTS corpus dari tantangan n2c2/OHNLP 2019 yang diselenggarakan oleh Harvard Medical School dan Mayo Clinic
	Metode	Model berbasis BERT yang ditingkatkan dengan tiga modul representasi: 1. Character-level (CNN) untuk mengatasi out-of-vocabulary 2. Sentence-level (BERT pretrained) untuk representasi semantik 3. Entity-level (Entity I & II) untuk menangkap konteks entitas medis berbasis MeSH
	Hasil	Model BERT dasar menghasilkan PCC = 0.848. Setelah menambahkan representasi karakter dan entitas, performa meningkat hingga PCC = 0.868, menunjukkan peningkatan signifikan dalam pemahaman semantik teks medis.
	Model Evaluasi	Pearson Correlation Coefficient (PCC)
	Tahun	2020 (Published on 29 December 2020, Vol 8 No 12) Preprint pertama: 10 August 2020, https://preprints.jmir.org/preprint/23357
8	Nama Penelitian	Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, & Rongrong Ji
	Judul Jurnal	BEAT: Bi-directional One-to-Many Embedding Alignment for Text-based Person Retrieval
	Jenis Data	Data sekunder (citra dan deskripsi teks)
	Sumber Data	Tiga dataset utama Text-based Person Retrieval: CUHK-PEDES, ICFG-PEDES, dan RSTPReID, serta tambahan uji pada MS-COCO, CUB, dan Flowers
	Metode	Model BEAT (Bi-directional One-to-Many Embedding Alignment) yang mengatasi masalah optimization direction pada pemetaan teks-gambar dengan menghasilkan banyak representasi fitur tanpa parameter tambahan.
	Hasil	Model mencapai state-of-the-art dengan hasil: CUHK-PEDES (R@1 = 65.61), ICFG-PEDES (R@1 = 58.25), RSTPReID (R@1 = 48.10).
	Model Evaluasi	Recall@K (R@1) pada berbagai dataset
	Tahun	2023
9	Nama Penelitian	Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax
	Penulis	Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, Ray Kurzweil
	Judul Jurnal	Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax

	Jenis Data	Data teks multibahasa (parallel corpus)
	Sumber Data	United Nations (UN) parallel corpus dan BUCC mining dataset
	Metode	Bi-directional dual-encoder dengan additive margin softmax untuk pembelajaran embedding kalimat multibahasa
	Hasil	Mencapai P@1 ≥ 86% pada semua bahasa, dan ~97% pada retrieval dokumen tingkat UN; unggul pada tugas BUCC mining
	Model Evaluasi	Precision at 1 (P@1), cosine similarity score
	Tahun	2019
10	Nama Penelitian	Multi-label Text Classification of Indonesian Customer Reviews using Bidirectional Encoder Representations from Transformers Language Model
	Penulis	Nuzulul Khairu Nissa & Evi Yulianti
	Judul Jurnal	International Journal of Electrical and Computer Engineering (IJECE), Vol. 13, No. 5, October 2023, pp. 5641–5652
	Jenis Data	Data teks ulasan pelanggan (customer review) berbahasa Indonesia
	Sumber Data	Dataset ulasan pelanggan (customer review) dari berbagai produk di platform daring Indonesia
	Metode	Dua strategi berbasis IndoBERT: (1) IndoBERT sebagai fitur dipadukan dengan CNN-XGBoost, (2) IndoBERT sebagai fitur sekaligus klasifikator utama; dibandingkan juga dengan multilingual BERT
	Hasil	Model IndoBERT meningkatkan efektivitas Word2Vec-CNN-XGBoost sebesar 19,19% (akurasi) dan 6,17% (F1-score); model kedua (IndoBERT sebagai klasifikator) paling unggul
	Model Evaluasi	Accuracy dan F1-score
	Tahun	2023
11	Nama Penelitian	Quasi Bidirectional Encoder Representations from Transformers (QBERT) for Word Sense Disambiguation
	Penulis	Michele Bevilacqua & Roberto Navigli
	Judul Jurnal	Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation
	Jenis Data	Data teks berbahasa Inggris (Word Sense Disambiguation corpus)
	Sumber Data	Dataset WSD berbasis WordNet dan korpus evaluasi WSD publik
	Metode	Arsitektur Transformer dengan lapisan co-attentive untuk menghasilkan representasi lebih mendalam (QBERT), dirancang khusus untuk tugas WSD
	Hasil	QBERT melampaui state of the art pada gabungan semua dataset WSD dengan peningkatan lebih dari 3 poin; mengungguli model berbasis ELMo
	Model Evaluasi	Akurasi dan perbandingan performa dengan ELMo
	Tahun	2020

Berdasarkan penelitian tersebut, dapat disimpulkan bahwa penerapan model BERT (Bidirectional Encoder Representations from Transformers) menunjukkan efektivitas tinggi di berbagai domain, mulai dari analisis sentimen, imputasi data hilang, hingga prediksi dan rekonstruksi teks. Penelitian Tri Buwono Bagus Wicaksono & Rama Dian Syah (2023) menunjukkan bahwa model IndoBERT mampu mengklasifikasikan sentimen pengguna aplikasi *Access by KAI* dengan akurasi tinggi (85%) dan penerapan praktis dalam bentuk prototipe website, meskipun masih terdapat kelemahan pada deteksi sentimen netral. Sementara itu, penelitian Llinet Benavides-Cesar dkk. (2025) membuktikan bahwa penerapan BERT untuk imputasi data spasio-temporal meningkatkan akurasi prediksi irradiance hingga 3%, dengan penurunan signifikan pada nilai MAE dan RMSE. Di sisi lain, penelitian Yusuke Mori dkk. (2020) memperluas penerapan BERT ke bidang *natural language generation* melalui model Missing Position Prediction (MPP) dan Story Completion (SC), yang mampu mengenali serta melengkapi bagian cerita yang hilang dengan akurasi 52,2%, bahkan menghasilkan narasi yang setara dengan teks asl pada 26% kasus. Secara keseluruhan, ketiga penelitian ini menegaskan bahwa BERT dan turunannya seperti IndoBERT, Sentence-BERT, serta integrasi BERT dengan arsitektur lain (GRU, CNN, dan Seq2Seq) mampu memberikan hasil yang unggul baik dalam pemahaman konteks bahasa maupun prediksi data kompleks, menjadikannya metode yang fleksibel dan berpotensi besar dalam berbagai bidang komputasi berbasis kecerdasan buatan.

2.2 Landasan Teori

Landasan teori merupakan kerangka konseptual yang menjadi dasar ilmiah penelitian dalam mengidentifikasi dan menyusun konsep serta variabel secara sistematis. Kerangka ini mengintegrasikan teori, metode, dan temuan penelitian sehingga proses analisis berlangsung secara logis, konsisten, dan dapat dipertanggungjawabkan secara akademik

2.2.1 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) merupakan model pra-latih (*pre-trained model*) yang dikembangkan oleh Devlin et al. (2019)

berdasarkan arsitektur *Transformer*. Berbeda dengan model berbasis arah tunggal seperti RNN atau LSTM, BERT mampu memahami konteks kata secara dua arah, baik dari kiri ke kanan maupun dari kanan ke kiri, sehingga menghasilkan representasi bahasa yang lebih mendalam dan kontekstual. Dengan pendekatan *bidirectional*, BERT dapat menangkap makna kata yang bergantung pada konteks sekitarnya secara simultan. Kemampuan ini menjadikan BERT unggul dalam berbagai tugas pemrosesan bahasa alami (*Natural Language Processing*), seperti analisis sentimen, klasifikasi teks, *question answering*, serta pelengkapan teks (*text completion*) (Wilkho et al., 2024).

Secara arsitektural, BERT dibangun sepenuhnya menggunakan komponen *encoder* dari *Transformer architecture* yang diperkenalkan oleh Vaswani et al. (2017). Arsitektur ini terdiri dari tumpukan beberapa lapisan *encoder* yang menggunakan mekanisme *self-attention* untuk menentukan hubungan antar kata dalam satu kalimat tanpa bergantung pada urutan posisi kata secara eksplisit. Mekanisme *self-attention* memungkinkan setiap token dalam kalimat untuk menimbang relevansi terhadap token lain, sehingga model dapat memahami konteks global dalam satu kali pemrosesan. Selain itu, BERT juga memanfaatkan *positional encoding* untuk memberikan informasi posisi pada setiap token agar model tetap mampu membedakan urutan kata dalam teks yang diproses (Satirapiwong & Siriborvornratanakul, 2021).

Dalam tahap pra-pelatihan (*pre-training*), BERT mengadopsi dua tujuan utama, yaitu *Masked Language Modeling (MLM)* dan *Next Sentence Prediction (NSP)*. Pada MLM, sebagian kata dalam kalimat ditutupi (*masked*) secara acak dan model diminta untuk memprediksi kata yang hilang berdasarkan konteks sekitarnya. Sementara itu, pada NSP, model belajar untuk memahami hubungan antar kalimat dengan menentukan apakah kalimat kedua merupakan lanjutan logis dari kalimat pertama atau tidak. Melalui dua mekanisme pelatihan ini, BERT memperoleh pemahaman kontekstual yang kuat, baik pada tingkat kata maupun antar kalimat, yang membuatnya mampu beradaptasi dengan berbagai tugas NLP melalui proses *fine-tuning* dengan data spesifik. Keunggulan ini menjadikan BERT sebagai fondasi penting dalam penelitian dan pengembangan sistem prediksi teks, termasuk dalam konteks melengkapi teks ulasan pengguna pada platform *airline review* (Satirapiwong & Siriborvornratanakul, 2021).

2.2.2 Machine Learning

Machine learning memanfaatkan metode untuk mengelola data dalam jumlah besar secara cerdas untuk menghasilkan output yang akurat. Berdasarkan metode pembelajaran yang digunakan, jenis-jenis machine learning dapat dibagi menjadi pembelajaran terawasi (supervised learning), pembelajaran tidak terawasi (unsupervised learning), pembelajaran semi terawasi (semi supervised learning), dan pembelajaran penguatan (reinforcement learning). Pembelajaran terawasi adalah salah satu metode machine learning yang mengandalkan dataset yang telah diberi label untuk melatih mesin, sehingga mesin dapat mengenali label input dengan memanfaatkan fitur yang ada untuk kemudian melakukan prediksi atau klasifikasi. Sementara itu, pembelajaran tidak terawasi adalah pendekatan yang berfokus pada penarikan kesimpulan berdasarkan dataset yang berisi input tanpa respon yang dilabeli (E.Retnoningsih C R.Pramudita, 2020).

Konsep utama dari machine learning dimulai pada tahun 1950-an saat Alan Turing memperkenalkan ide mengenai "Turing Test" untuk mengukur kecerdasan mesin (Turing, 1950). Di tahun 1952, Arthur Samuel menciptakan sebuah program komputer yang bisa belajar untuk bermain permainan dam (checkers), yang sering dianggap sebagai salah satu contoh awal dari machine learning (Samuel, 1959) . Sejalan dengan pertumbuhan data digital dan peningkatan daya komputasi, ML mengalami perkembangan yang signifikan. Ide Deep Learning menggunakan jaringan syaraf tiruan bertingkat telah dikembangkan lebih lanjut oleh Hinton dan rekan-rekannya pada tahun 2006. Sampai sekarang, ML diterapkan di berbagai sektor, termasuk pengenalan suara, pengolahan bahasa alami (NLP), analisis media sosial, serta pemilu digital. Dengan kemampuan menganalisis data besar secara efisien serta menghasilkan prediksi akurat, Machine Learning terus membuka peluang baru di berbagai sektor industri.

2.2.3 Text Preprocesing

Pengolahan awal data (*data preprocessing*) merupakan tahap krusial dalam penelitian berbasis analisis teks, khususnya pada pemrosesan bahasa alami (*Natural Language Processing*). Tahap ini bertujuan untuk membersihkan data teks mentah dari elemen-elemen yang tidak relevan sehingga dapat meningkatkan kualitas representasi data sebelum digunakan dalam proses pemodelan. Data yang diperoleh melalui teknik *web scraping* umumnya masih mengandung noise seperti tautan URL, simbol, serta variasi penulisan yang tidak konsisten. Sebagaimana dijelaskan dalam penelitian Djunaidi et al. (2025), tahap *text preprocessing* berperan penting dalam menyelaraskan dan meningkatkan kualitas data agar siap digunakan pada proses analisis lanjutan(Review_of_Data_Preprocessing_Techniques.pdf n.d.). Dalam penelitian ini, tahapan preprocessing dilakukan secara bertahap meliputi *case folding* (mengubah seluruh teks menjadi huruf kecil), penghapusan URL, penghapusan karakter non-alfanumerik, serta normalisasi spasi untuk menghilangkan spasi berlebih dalam teks. Tahapan ini bertujuan untuk menghasilkan struktur teks yang lebih konsisten dan bersih sebelum dilakukan proses tokenisasi menggunakan BERT Tokenizer. Dengan melakukan pembersihan teks secara sistematis, model dapat mempelajari pola bahasa secara lebih optimal tanpa terganggu oleh elemen-elemen yang tidak memiliki kontribusi terhadap makna kalimat.

2.2.4 Web Scraping

Web scraping merupakan teknik pengumpulan data secara otomatis dari halaman website dengan cara mengekstraksi informasi yang tersedia pada struktur HTML suatu situs. Teknik ini banyak digunakan dalam penelitian berbasis analisis teks dan data mining untuk memperoleh dataset dari sumber daring yang bersifat publik.

Berdasarkan penelitian Siswipraptini et al. (2023), web scraping digunakan sebagai tahap awal dalam proses pengumpulan data dari website lowongan pekerjaan untuk selanjutnya dilakukan proses *text preprocessing* dan klasifikasi dokumen (Siswipraptini, Wafit, and Djunaidi 2023). Dalam penelitian tersebut dijelaskan bahwa hasil dari web scraping berupa data mentah yang masih mengandung berbagai elemen seperti deskripsi teks, lokasi, dan informasi lainnya, sehingga memerlukan tahap pembersihan sebelum dilakukan pemodelan.

Secara umum, proses web scraping melibatkan beberapa langkah, yaitu mengakses halaman web, membaca struktur dokumen HTML, mengekstraksi elemen yang dibutuhkan, serta menyimpan data dalam format terstruktur seperti CSV atau database. Teknik ini memungkinkan peneliti memperoleh data dalam jumlah besar secara efisien tanpa melakukan pengumpulan data secara manual.

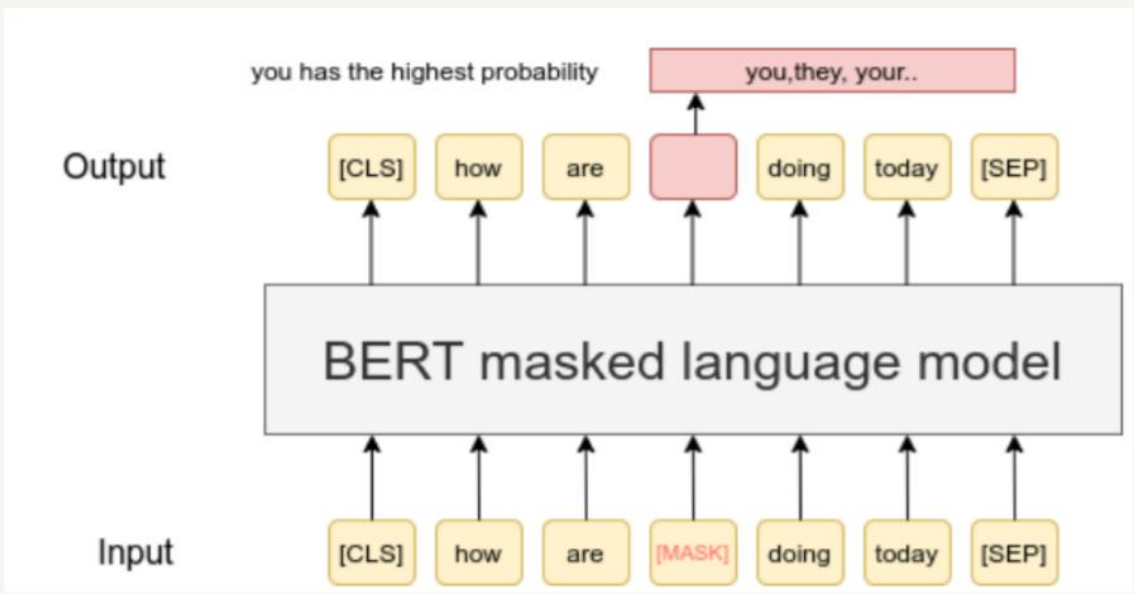
Dalam penelitian ini, teknik web scraping digunakan untuk mengumpulkan data ulasan pengguna maskapai penerbangan dari platform Airline Review (Skytrax). Data yang diperoleh berupa teks ulasan yang kemudian melalui tahap preprocessing sebelum dilakukan proses tokenisasi dan fine-tuning model BERT untuk memprediksi missing text.

2.2.5 Python

Python adalah bahasa pemrograman tingkat tinggi yang berorientasi objek dan dibuat oleh Guido van Rossum. Python adalah bahasa lintas fungsi yang ditafsirkan secara maksimal yang memiliki banyak fungsi. Bahasa pemrograman berorientasi objek ini biasanya digunakan untuk merampingkan kumpulan data kompleks yang besar. Fitur lain yang membuat Python ini terkenal di kalangan ilmuwan dan penganalisis data adalah fleksibilitasnya. Karena fleksibilitasnya, bahasa ini memungkinkan untuk membuat data model, mensistemasikan kumpulan data, membuat algoritma yang didukung ML (Machine Learning), layanan web, dan menerapkan penambangan data untuk menyelesaikan berbagai tugas dalam waktu singkat (Junaidi et al., 2023) . Python memiliki perkembangan popularitas yang luar biasa dalam komunitas scientific computing. Hal ini dikarenakan banyaknya library machine learning dan deep learning menggunakan python sebagai bahasa utama mereka. Beberapa contoh library yang menerapkan machine learning adalah Bidirectional Encoder Representations from Transformer atau BERT yang dikenal sebagai dasar dari beberapa library yang digunakan untuk menganalisis data seperty PyABSA dan BERTopic (Listyawan et al., 2024) .

2.2.6 Masked Language Modeling (MLM)

Masked Language Modeling (MLM) adalah salah satu konsep utama yang digunakan dalam tahap pelatihan awal model BERT (Bidirectional Encoder Representations from Transformers). Ide dasarnya cukup sederhana:sebagian kata dalam sebuah kalimat sengaja disembunyikan, lalu model diminta untuk menebak kata yang hilang tersebut dengan melihat konteks dari kata-kata lain di sekitarnya (Devlin et al., 2019). Dengan cara ini, BERT tidak hanya belajar mengenali urutan kata, tetapi juga memahami makna dan hubungan antarkata secara lebih menyeluruh. Pendekatan ini membuat model mampu memahami konteks kalimat secara dua arah dari kiri ke kanan dan dari kanan ke kiri sehingga pemahaman yang dihasilkan lebih natural dan akurat (Satirapiwong & Siriborvornratanakul, 2021). Berikut dibawah ini saya tampilkan arsitektur BERT-MLM nya pada gambar 2.1



Gambar2. 1 Arsitektur BERT-MLM

Tujuan utama dari MLM adalah untuk memprediksi kata yang hilang (*missing text*) dalam sebuah kalimat. Sebagian token (biasanya 15%) dalam input kalimat di-*masking*, dan model dilatih untuk memprediksi token yang benar berdasarkan konteks sekitarnya. *Masked Language Modeling* (MLM) secara matematis dapat dirumuskan menggunakan Persamaan 1.

$$L_{MLM} = -\sum_{i \in M} \log P(\omega_i \mid \omega_{\setminus M}) \quad (1)$$

Berdasarkan Persamaan 1, M menyatakan himpunan indeks token yang di-*mask* dalam suatu urutan teks. Simbol ω_i merepresentasikan token asli pada posisi ke-i yang disembunyikan dan dijadikan sebagai target prediksi. Notasi $\omega_{\setminus M}$ menunjukkan seluruh token dalam kalimat yang tidak termasuk dalam himpunan *mask*, sehingga berfungsi sebagai konteks dalam proses prediksi. Sementara itu, $P(\omega_i \mid \omega_{\setminus M})$ menyatakan probabilitas kemunculan token asli yang diprediksi oleh model berdasarkan informasi konteks yang tersedia. Fungsi kerugian ini bertujuan untuk meminimalkan nilai negatif log-likelihood dari token-token yang di-*mask*, sehingga model terdorong untuk menghasilkan prediksi yang semakin mendekati distribusi kata sebenarnya.

Dalam praktiknya, sekitar 15% kata dalam kalimat diganti dengan token khusus [MASK] selama proses pelatihan. Model kemudian berusaha memprediksi kata yang tersembunyi itu berdasarkan konteks yang tersedia. Misalnya, pada kalimat "Penumpang merasa [MASK] dengan pelayanan maskapai," model BERT dapat memperkirakan bahwa kata yang hilang kemungkinan besar adalah "puas." Proses ini dilakukan menggunakan mekanisme self-attention, di mana setiap kata dalam kalimat memperhatikan kata lainnya untuk membangun pemahaman makna secara menyeluruh. Hasilnya, model mampu mengenali pola bahasa dengan presisi tinggi dan menyesuaikan makna kata sesuai dengan konteks kalimatnya (Pan et al., 2021).

Dalam penelitian mengenai user experience di platform airline review, konsep MLM sangat berguna untuk melengkapi atau memperbaiki teks ulasan pengguna yang tidak lengkap. Sering kali, ulasan pelanggan di internet berisi kalimat yang terpotong, kata yang hilang, atau penulisan yang tidak sempurna. Dengan pendekatan MLM, sistem dapat "menebak" kata yang hilang secara alami sesuai konteks kalimat, sehingga menghasilkan ulasan yang lebih lengkap dan mudah dipahami (Cunha Sergio & Lee, 2021). Pendekatan ini telah terbukti efektif pada penelitian Riccardi dan Bandara (2024), yang menggunakan model DistilBERT untuk memprediksi kata hilang pada teks kuno berbahasa Yunani dan berhasil mencapai tingkat akurasi yang tinggi. Hal ini menunjukkan bahwa teknik MLM dapat diadaptasi dengan baik untuk memperbaiki data teks modern, termasuk ulasan pengguna dalam konteks pengalaman penerbangan.

2.2.7 ROUGE -1

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) adalah serangkaian metrik evaluasi otomatis yang digunakan untuk menilai kualitas teks yang dihasilkan oleh suatu sistem dibandingkan dengan teks referensi manusia. ROUGE pertama kali diperkenalkan oleh *Chin-Yew Lin* sebagai alat evaluasi dalam tugas ringkasan teks otomatis (*automatic text summarization*), dan sejak itu telah digunakan luas dalam berbagai tugas Natural Language Processing (NLP) untuk mengevaluasi hasil generatif teks yang dihasilkan model terhadap teks referensi.

Dalam konteks ROUGE, yang dibandingkan adalah tumpang tindih (overlap) antara unit-unit n-gram (urutan kata) yang muncul dalam teks keluaran yang dihasilkan oleh model dan teks referensi. Ukuran overlap ini kemudian dapat dihitung dalam bentuk *recall*, *precision*, atau *F1 score* untuk mendapatkan angka evaluasi yang merepresentasikan tingkat kesamaan antara dua teks.

ROUGE-1 adalah varian ROUGE yang mengukur overlap unigram saja, yaitu kesamaan pada tingkat kata tunggal antara teks hasil prediksi model dan teks referensi. ROUGE-1 dihitung dengan membandingkan jumlah unigram yang terdapat pada teks hasil prediksi dengan unigram yang sama pada teks referensi. *Rouge-1* secara matematis dapat dirumuskan menggunakan Persamaan 2.

$$ROUGE-1 = \frac{\sum \omega \epsilon Reference \; \min(Count_{pred(\omega)}, Count_{ref(\omega)})}{\sum_{\omega \epsilon Reference} Count_{ref(\omega)}} \quad (2)$$

Berdasarkan Persamaan 2, symbol ω merepresentasikan satuan kata tunggal (*unigram*) yang terdapat dalam teks referensi. Notasi $Count_{pred(\omega)}$ menyatakan jumlah kemunculan kata ω pada teks hasil prediksi model, sedangkan $Count_{ref(\omega)}$ menunjukkan jumlah kemunculan kata yang sama pada teks referensi (teks asli). Fungsi minimum min $Count_{pred(\omega)}$, $Count_{ref(\omega)}$ digunakan untuk menghitung jumlah unigram yang benar-benar tumpang tindih (*overlap*) antara teks prediksi dan teks referensi.

kemudian dinormalisasi relatif terhadap jumlah total unigram pada teks referensi.

Skor ROUGE-1 yang lebih tinggi menunjukkan bahwa model mampu menghasilkan kata-kata penting yang sama seperti yang terdapat pada teks referensi, sehingga mencerminkan kesesuaian konten secara leksikal.

ROUGE-1 dipilih dalam penelitian ini karena fokus evaluasi berada pada prediksi kata atau frasa pendek yang hilang dalam ulasan pengguna, di mana keterkaitan konteks dan relevansi kata sangat penting. Evaluasi berdasarkan unigram memungkinkan pengukuran seberapa besar model mempertahankan kata-kata penting dari teks referensi dalam hasil prediksi yang dihasilkan(Tay et al. 2004).

Dalam banyak penelitian terkait evaluasi ringkasan teks dan teks generatif, ROUGE-1 menjadi salah satu metrik standar karena mampu memberikan gambaran objektif tentang tingkat kesamaan konten antara teks hasil model dan teks referensi manusia.

2.2.8 CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan kerangka kerja standar yang banyak digunakan dalam pelaksanaan proyek *data mining* di berbagai sektor industri(Rifai, Jatnika, and Valentino 2019). Model ini menyediakan pendekatan yang sistematis dan terstruktur melalui enam tahapan utama, sehingga proses analisis data dapat dilaksanakan secara terarah, terintegrasi, dan selaras dengan tujuan strategis organisasi.

Tahap pertama yaitu *business understanding*, yang menitikberatkan pada pemahaman komprehensif terhadap tujuan, kebutuhan, serta permasalahan bisnis yang mendasari proyek. Pada fase ini dilakukan perumusan masalah, penetapan sasaran, serta penyusunan rencana proyek yang relevan dengan kepentingan organisasi. Wirth dan Hipp (2000) menyatakan bahwa tahap ini berfokus pada penerjemahan tujuan dan kebutuhan bisnis ke dalam definisi permasalahan *data mining* serta perancangan rencana proyek awal yang dirumuskan untuk mencapai tujuan tersebut. Dengan demikian, tahap ini menjadi landasan konseptual bagi tahapan selanjutnya.

Setelah permasalahan terdefinisi dengan jelas, proses berlanjut ke tahap *data understanding*. Tahap ini meliputi pengumpulan data awal, eksplorasi karakteristik dan struktur data, identifikasi permasalahan kualitas data, serta analisis awal guna

memperoleh pemahaman terhadap pola, anomali, atau informasi yang terkandung di dalam data. Pemahaman yang mendalam terhadap data menjadi prasyarat penting sebelum dilakukan pengolahan lebih lanjut.

Selanjutnya, tahap *data preparation* mencakup proses pembersihan, transformasi, integrasi, dan seleksi fitur untuk memastikan data berada dalam kondisi yang optimal untuk pemodelan. Tahap ini bertujuan menjamin kualitas, konsistensi, dan kesesuaian struktur data sehingga mampu mendukung pembentukan model yang valid dan akurat (Review_of_Data_Preprocessing_Techniques.pdf n.d.). Keberhasilan tahap ini sangat menentukan efektivitas proses pemodelan yang akan dilakukan.

Tahap *modeling* kemudian dilaksanakan dengan memilih metode atau algoritma yang sesuai, membagi data ke dalam data latih dan data uji, membangun model, serta mengevaluasi performanya. Proses ini umumnya bersifat iteratif guna memperoleh model dengan tingkat kinerja yang optimal (Goud et al., 2024).

Setelah model terbentuk, dilakukan tahap *evaluation* untuk menilai kinerja model secara menyeluruh serta memastikan kesesuaiannya dengan tujuan bisnis yang telah ditetapkan. Pada fase ini juga dilakukan peninjauan ulang terhadap keseluruhan proses guna mengidentifikasi potensi kelemahan sebelum model dinyatakan layak untuk diimplementasikan.

Tahap terakhir adalah *deployment*, yaitu proses penerapan model ke dalam lingkungan operasional, baik dalam bentuk penyusunan laporan, integrasi ke dalam sistem, maupun pengembangan mekanisme pemantauan berkelanjutan. Melalui tahap ini, hasil analisis tidak hanya berhenti pada aspek teoretis, tetapi diimplementasikan secara nyata untuk mendukung pengambilan keputusan strategis oleh para pemangku kepentingan.

2.2.9 *User Experience (UX)*

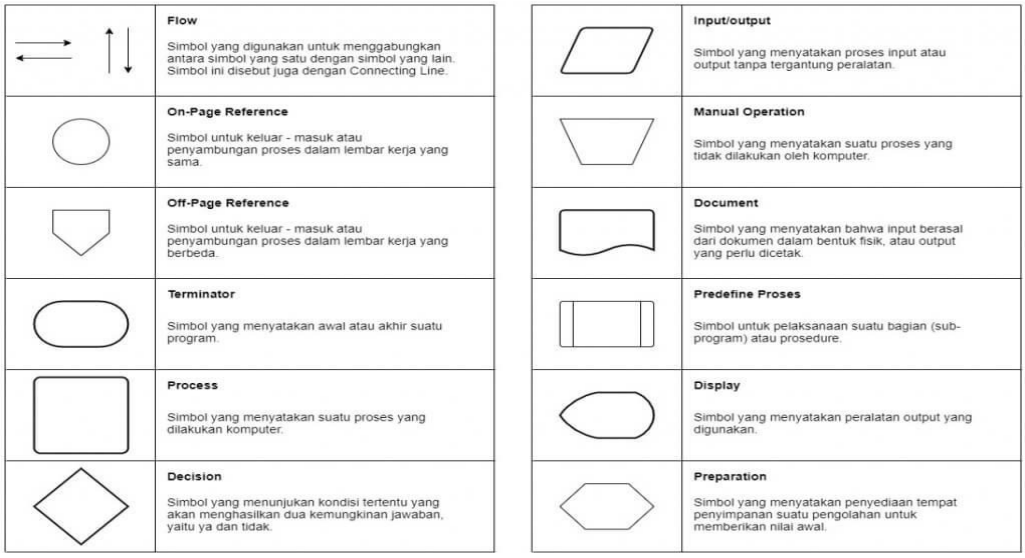
User Experience (UX) merupakan konsep yang menggambarkan seluruh pengalaman dan persepsi pengguna ketika berinteraksi dengan suatu produk atau layanan, baik itu aplikasi digital, situs web, maupun sistem berbasis teknologi. Menurut standar ISO 9241-210, UX mencakup aspek emosional, kognitif, dan perilaku pengguna selama proses penggunaan produk. Dengan kata lain, UX tidak hanya menilai seberapa baik suatu

sistem berfungsi, tetapi juga bagaimana sistem tersebut membuat pengguna merasa nyaman, puas, dan mudah dalam mencapai tujuannya. Dalam konteks digital, UX menjadi kunci utama dalam menentukan keberhasilan sebuah platform karena pengalaman yang positif dapat meningkatkan tingkat kepuasan, loyalitas, dan kepercayaan pengguna terhadap layanan (Pan et al., 2021).

Dalam penelitian yang berkaitan dengan *airline review platform*, UX memiliki peran penting karena ulasan yang diberikan oleh pengguna mencerminkan pengalaman mereka secara langsung terhadap layanan maskapai penerbangan, seperti kenyamanan, ketepatan waktu, kualitas pelayanan, dan fasilitas yang diberikan. Ulasan ini menjadi sumber informasi berharga bagi perusahaan untuk memahami persepsi pelanggan dan melakukan perbaikan layanan. Namun, tantangan muncul ketika ulasan pengguna tidak lengkap atau terdapat teks yang hilang, sehingga makna pengalaman tidak tersampaikan secara utuh. Oleh karena itu, pendekatan berbasis *Natural Language Processing (NLP)* seperti BERT dan *Masked Language Modeling (MLM)* dapat dimanfaatkan untuk melengkapi atau memperkirakan bagian teks yang hilang. Dengan cara ini, analisis UX menjadi lebih akurat dan representatif terhadap pengalaman nyata pengguna (Cunha Sergio & Lee, 2021).

2.2.10 Flowchart

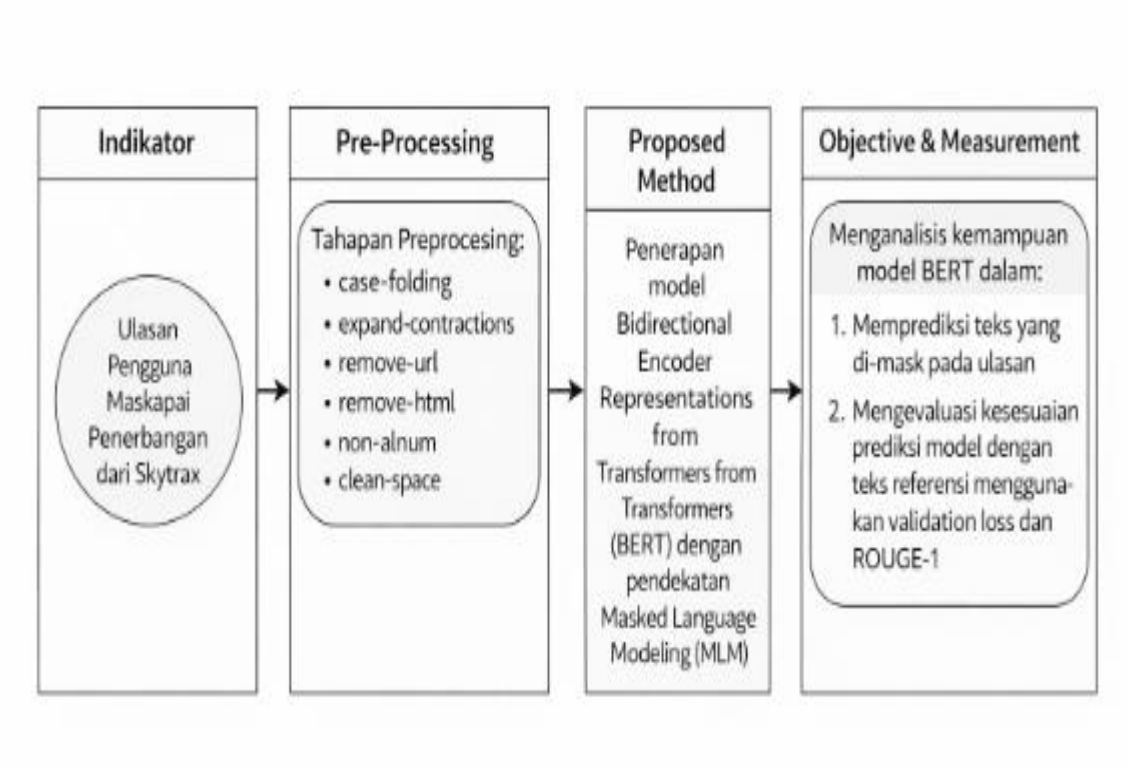
Flowchart, atau sering disebut sebagai diagram alir, adalah representasi visual dari suatu algoritma atau serangkaian langkah sistematis dalam sebuah sistem. Diagram ini digunakan oleh analis sistem untuk mendokumentasikan serta menjelaskan gambaran logis dari suatu sistem yang akan dikembangkan kepada programmer. Dengan adanya flowchart, proses pemecahan masalah dalam pengembangan sistem dapat lebih mudah dipahami dan diimplementasikan (Rosaly et al., n.d.). Flowchart terdiri dari berbagai simbol standar yang masing-masing memiliki fungsi tertentu. Simbol-simbol ini digunakan untuk menggambarkan proses, keputusan, input, output, dan arah alur kerja dalam sistem. Garis penghubung digunakan untuk menghubungkan satu proses ke proses lainnya sehingga dapat membentuk urutan logis yang jelas (Shelly & Rosenblatt, 2012).



Gambar2. 2 Simbol symbol flowchart

(Ridho, 2024)

2.3 Kerangka Pemikiran



Gambar2. 3 Kerangka Pemikiran

Gambar 2.3 menunjukkan alur penelitian yang digunakan dalam penerapan model BERT dengan pendekatan Masked Language Modeling (MLM) pada data ulasan pengguna maskapai penerbangan. Tahapan penelitian dimulai dari pengumpulan data ulasan dari platform Airline Review (Skytrax), kemudian dilakukan proses *preprocessing* yang meliputi pembersihan teks dan normalisasi data.Selanjutnya, diterapkan model BERT dengan pendekatan MLM, di mana sebagian token dalam teks di-mask untuk mensimulasikan teks yang tidak lengkap. Model kemudian dilatih untuk memprediksi token tersebut berdasarkan konteks kalimat. Evaluasi performa model dilakukan menggunakan *validation loss* dan metrik ROUGE-1 untuk mengukur tingkat kesesuaian antara hasil prediksi dan teks referensi.