BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Berikut merupakan penelitian terdahulu yang dapat mendukung dilakukannya penelitian ini:

Tabel 2.1 Penelitian Terdahulu

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
1	Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision – V. D. Setiawan et al. (2024) [23]	Scientific Journal of Informatics	IndoBERT, BERT, SVM, Random Forest	IndoBERT mencapai akurasi tertinggi 95%, mengungguli BERT (92%), SVM (88%), dan Random Forest (85%).
2	Perbandingan Kinerja IndoBERT dan MBERT untuk Deteksi Berita Hoaks Politik – C. J. L. Tobing et al. (2024) [24]	Jurnal Sains dan Teknologi	IndoBERT vs MBERT	IndoBERT memperoleh akurasi 95,6% dengan F1-score 93,6%, lebih tinggi dibanding MBERT (91,97%).
3	Perbandingan IndoBERT dan IndoRoBERTa Untuk Analisis Sentimen Pada Film Dokumenter Dirty Vote – F. M. Apriansyah et al. (2024) [25]	Jurnal Informatika: Jurnal Pengembangan IT	IndoBERT vs IndoRoBERTa	IndoBERT mencapai akurasi 99%, sedangkan IndoRoBERTa 94%.
4	Penerapan IndoBERT dan BERTopic dalam ABSA untuk Evaluasi Kualitas Aplikasi E-Government – N. A.	Jurnal Teknologi dan Sistem Informasi Univrab	BERTopic + IndoBERT + SVM + Naive Bayes	IndoBERT menunjukkan performa terbaik dengan akurasi 91%, melampaui SVM (84%) dan Naive Bayes (80%). Namun, proses pelabelan data

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
	Adrielvino et al. (2024) [26]			ulasan masih mengandalkan annotator manusia secara manual dan penelitian belum diimplementasikan ke dalam sistem produksi atau prototipe.
5	Aspect-Based Sentiment Analysis of Access by KAI Application Reviews Using IndoBERT – H. N. Alfiana et al. (2024) [27]	Jurnal Teknik Informatika (JUTIF)	IndoBERT vs XLM-R, SVM, XGBoost, BiLSTM, mBERT	IndoBERT mencapai akurasi 96,2% pada klasifikasi gabungan aspek-sentimen, sementara XLM-R menunjukkan performa kompetitif dengan akurasi 96,1%, 92,5% pada klasifikasi aspek, dan 93,2% pada klasifikasi sentimen. Meskipun membandingkan kedua model tersebut, penentuan aspek ulasan di dalam studi ini masih ditentukan secara manual (bukan otomatis) dan belum mencakup aspek visualisasi deployment.
6	Fine Tuning IndoBERT Untuk Analisis Sentimen Pada Ulasan Pengguna Aplikasi Tiket.com – D. Nuryadi et al. (2024) [28]	JATI	Fine-tuned IndoBERT	Model mencapai akurasi 91%, dengan F1-score 93% (positif) dan 94% (negatif).
7	Klasifikasi Sentimen Google Play Store Aplikasi ChatGPT Berbahasa Indonesia Berbasis IndoBERT – I.	Jurnal Minfo Polgan	IndoBERT	IndoBERT mencapai akurasi 89%, precision 87%, recall 89%, dan F1-score 88%. Namun, analisis sentimen yang dilakukan masih berada pada level dokumen

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
	Mursidah et al. (2024)[6]			(secara umum), belum membedakan sentimen berdasarkan aspek-aspek spesifik aplikasi.
8	Fine-Tuned IndoBERT for Aspect-Based Sentiment Analysis of Indonesian Five-Star Hotel Reviews – S. Apriliani et al. (2024)[29]	Jurnal Sisfokom (Sistem Informasi dan Komputer)	Fine-tuned IndoBERT ABSA	Model mencapai akurasi 95,28% dengan Macro F1-score 82,44%.
9	ABSA of Indonesian Customer Reviews Using IndoBERT: Single-Sentence and Sentence-Pair Classification – E. Yulianti et al. (2024) [30]	Bulletin of Electrical Engineering and Informatics	IndoBERT Sentence-Pair vs Single-Sentence	Pendekatan sentence-pair menghasilkan performa terbaik dengan F1-score 0,9194.
10	Aspect-Based Sentiment Analysis of App Reviews for Requirement Elicitation – S. Taj et al. (2024) [31]	Engineering Applications of Artificial Intelligence	Fine-tuned BERT + XAI	Model mencapai F1-score 0,83 untuk Aspect Category Detection (ACD) dan 0,94 untuk Aspect Category Polarity (ACP).
11	Analisis Sentimen Ulasan Aplikasi ChatGPT Pada Google Play Menggunakan Metode SVM – A. Salsabila et al. (2024) [8]	Jurnal Ilmiah Komputer	Support Vector Machine	Model SVM mencapai akurasi 91%.
12	Analisis Sentimen Ulasan Aplikasi Fintech menggunakan Framework CRISP-DM	Jurnal Sistem Informasi	SVM + CRISP-DM	Framework CRISP-DM menghasilkan akurasi 86%.

No	Judul dan Peneliti	Nama Jurnal	Metode	Hasil Penelitian
	– M. R. Amalsyah et al. (2024) [32]			
13	Sentiment Analysis for Low-Resource Languages Using a Hybrid XLM-RoBERTa and Bi-LSTM Model – Mohan. B et al (2025) [33]	*2025 International Conference on Networks & Advances in Computational Technologies* (NetACT)Evaluation Conference	Hybrid XLM-RoBERTa dan Bi-LSTM	Model hybrid XLM-RoBERTa dan Bi-LSTM digunakan untuk analisis sentimen pada bahasa *low-resource* dan teks *code-mixed* dengan akurasi 62%. Penelitian ini relevan sebagai pendukung penggunaan XLM-RoBERTa dalam klasifikasi sentimen multibahasa.

Berdasarkan Tabel 2.1, dapat diketahui bahwa penelitian terkait analisis sentimen dan Aspect-Based Sentiment Analysis (ABSA) mengalami perkembangan yang signifikan, khususnya pada penerapan model berbasis Transformer untuk teks berbahasa Indonesia. Berbagai penelitian menunjukkan bahwa model monolingual seperti IndoBERT memiliki performa yang sangat kompetitif dibandingkan metode machine learning tradisional maupun model Transformer lainnya. Penelitian sebelumnya menunjukkan bahwa Model IndoBERT terbukti mampu menghasilkan performa yang baik pada berbagai tugas klasifikasi teks berbahasa Indonesia, baik pada domain hukum, berita, maupun ulasan aplikasi [23], [24].

Dalam konteks analisis sentimen pada domain ulasan aplikasi, IndoBERT juga menunjukkan performa yang konsisten. Penelitian sebelumnya pada ulasan aplikasi Tiket.com menunjukkan akurasi sebesar 91%, sementara penelitian oleh Mursidah et al. pada ulasan aplikasi ChatGPT menunjukkan akurasi sebesar 89% [6], [28]. Hasil ini menunjukkan bahwa IndoBERT memiliki kemampuan representasi semantik yang kuat dalam memahami teks ulasan berbahasa Indonesia yang cenderung informal dan kontekstual.

Pada ranah ABSA, beberapa penelitian menunjukkan bahwa pendekatan berbasis Transformer semakin banyak digunakan untuk meningkatkan akurasi analisis. Penelitian sebelumnya menunjukkan bahwa kombinasi BERTopic dan IndoBERT menghasilkan performa lebih baik dibandingkan metode klasik seperti SVM dan Naive Bayes [26]. Penelitian lain juga menunjukkan bahwa IndoBERT maupun XLM-R mampu memberikan performa yang kompetitif dalam tugas klasifikasi aspek dan sentimen, dengan akurasi di atas 96% pada skenario joint classification [27]. Selain itu, penelitian sebelumnya menunjukkan bahwa IndoBERT efektif digunakan dalam ABSA berbahasa Indonesia, termasuk pada pendekatan sentence-pair classification yang lebih sesuai untuk menangkap hubungan antara aspek dan polaritas sentiment [29], [34].

Selain model monolingual, model multilingual juga menunjukkan potensi dalam tugas analisis sentimen. Penelitian sebelumnya menunjukkan bahwa XLM-RoBERTa dapat digunakan untuk analisis sentimen pada teks multibahasa dan *low-resource* dengan memanfaatkan kemampuan representasi kontekstual lintas bahasa [33]. Temuan tersebut memperkuat bahwa model multilingual seperti XLM-RoBERTa relevan untuk dibandingkan dengan model monolingual seperti IndoBERT, khususnya pada domain ulasan aplikasi yang dapat mengandung variasi bahasa informal maupun campuran.

Dari sisi metodologi, beberapa penelitian juga menunjukkan bahwa framework CRISP-DM masih relevan digunakan dalam penelitian analisis sentimen karena memberikan alur kerja yang sistematis mulai dari pengumpulan data hingga deployment model. Penelitian sebelumnya menunjukkan implementasi CRISP-DM yang efektif pada domain analisis sentimen fintech, sehingga pendekatan ini menjadi landasan metodologis yang sesuai untuk penelitian ini[32].

Berdasarkan telaah pustaka terhadap tiga belas penelitian terdahulu yang disajikan pada Tabel 2.1, terdapat beberapa catatan penting yang menjadi landasan sekaligus pembeda bagi penelitian ini. Penelitian yang menganalisis ulasan ChatGPT sejauh ini masih terbatas pada analisis sentimen global tingkat dokumen, sehingga belum mampu menangkap keluhan atau apresiasi pengguna secara

granular. Di sisi lain, penelitian yang telah menerapkan *Aspect-Based Sentiment Analysis* (ABSA) menggunakan arsitektur *transformer*, umumnya masih menghadapi keterbatasan pada proses penentuan aspek yang dilakukan secara manual, atau proses anotasi data yang membutuhkan waktu lama karena bergantung sepenuhnya pada *human annotator*[26], [27], [29]. Selain itu, pengujian komparatif yang mempertemukan model spesifik bahasa seperti IndoBERT dengan model *multilingual* seperti XLM-RoBERTa dalam satu skenario ABSA berbahasa Indonesia masih sangat jarang dilakukan.

Penelitian ini dirancang secara khusus untuk mengatasi batasan-batasan dari studi terdahulu tersebut melalui pendekatan metodologi yang lebih komprehensif pada domain aplikasi AI generatif. Perbedaan mendasar sekaligus nilai kebaruan (*novelty*) dari penelitian ini terletak pada integrasi penuh empat komponen utama, yakni ekstraksi aspek secara otomatis menggunakan *Latent Dirichlet Allocation* (LDA), efisiensi pelabelan dataset skala besar berbasis *Large Language Model* (Qwen), pengujian komparatif performa model IndoBERT dan XLM-RoBERTa, serta penyediaan luaran praktis berupa analisis kata opini (*opinion words*) dan simulasi *deployment* prototipe sistem interaktif.

Oleh karena itu, penelitian ini bertujuan untuk mengisi gap tersebut dengan mengusulkan pendekatan ABSA terintegrasi yang mencakup ekstraksi aspek menggunakan topic modeling, pelabelan sentimen otomatis berbasis LLM, komparasi performa IndoBERT dan XLM-RoBERTa, serta implementasi prototype deployment untuk simulasi penggunaan model pada skenario nyata.

2.2 Teori yang berkaitan

2.2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP memungkinkan komputer untuk memahami, menganalisis, serta menghasilkan teks dalam bahasa manusia. Berbagai tugas NLP meliputi pemrosesan teks, klasifikasi teks, ekstraksi informasi, serta analisis sentimen[1].

Pada penelitian ini, NLP digunakan untuk memproses teks ulasan pengguna aplikasi ChatGPT yang ditulis dalam bahasa Indonesia. Proses NLP biasanya melibatkan beberapa tahapan seperti tokenisasi, pembersihan teks, normalisasi kata, serta transformasi teks menjadi representasi numerik yang dapat diproses oleh model machine learning atau deep learning[35]. Dengan menggunakan teknik NLP, opini pengguna dalam bentuk teks dapat dianalisis untuk mengetahui sentimen yang terkandung dalam ulasan tersebut.

2.2.2 Analisis Sentimen (Sentiment Analysis)

Analisis sentimen merupakan teknik dalam NLP yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini atau perasaan yang terkandung dalam suatu teks. Tujuan utama analisis sentimen adalah untuk menentukan apakah suatu opini bersifat positif, negatif, atau netral[36].

Dalam konteks ulasan aplikasi, analisis sentimen dapat digunakan untuk memahami persepsi pengguna terhadap suatu aplikasi berdasarkan komentar yang mereka berikan. Dengan menganalisis sentimen pada ulasan pengguna, pengembang aplikasi dapat mengetahui aspek-aspek yang mempengaruhi kepuasan maupun ketidakpuasan pengguna terhadap aplikasi yang mereka gunakan[32].

2.2.3 Aspek dalam Analisis Sentimen

Dalam analisis sentimen, aspek merujuk pada atribut, fitur, atau komponen spesifik dari suatu produk atau layanan yang menjadi fokus pembahasan dalam suatu teks ulasan. Aspek mencerminkan bagian tertentu yang dinilai oleh pengguna, seperti kualitas layanan, performa sistem, fitur aplikasi, kemudahan penggunaan, maupun karakteristik lain yang relevan dengan pengalaman pengguna[18].

Keberadaan aspek menjadi penting karena dalam satu ulasan, pengguna tidak selalu menyampaikan opini secara umum, melainkan dapat memberikan penilaian terhadap beberapa aspek yang berbeda secara bersamaan[37]. Sebagai contoh, dalam satu ulasan pengguna dapat menyatakan bahwa suatu aplikasi memiliki fitur yang sangat membantu, namun di sisi lain memiliki performa yang kurang stabil. Dalam kasus seperti ini, analisis sentimen konvensional yang hanya

menghasilkan satu label sentimen tidak mampu menangkap perbedaan opini terhadap masing-masing aspek tersebut.

Oleh karena itu, identifikasi aspek dalam analisis sentimen memungkinkan proses analisis dilakukan secara lebih rinci dan terstruktur. Dengan mengetahui aspek yang dibahas, analisis tidak hanya berfokus pada polaritas sentimen, tetapi juga mampu mengungkap faktor-faktor spesifik yang memengaruhi persepsi pengguna[29]. Hal ini memberikan nilai tambah dalam memahami kebutuhan pengguna serta menjadi dasar dalam pengambilan keputusan, khususnya dalam pengembangan dan evaluasi suatu sistem atau aplikasi.

Dalam konteks penelitian ini, aspek digunakan sebagai dasar untuk mengelompokkan ulasan pengguna aplikasi ChatGPT berdasarkan topik yang dibahas. Aspek-aspek tersebut tidak ditentukan secara manual, melainkan diidentifikasi secara otomatis menggunakan pendekatan berbasis data melalui metode topic modeling. Dengan demikian, aspek yang dihasilkan diharapkan mampu merepresentasikan pola utama dalam ulasan pengguna secara lebih objektif.

2.2.4 *Aspect-Based Sentiment Analysis* (ABSA)

Aspect-Based Sentiment Analysis (ABSA) merupakan pengembangan dari metode analisis sentimen yang bertujuan untuk mengidentifikasi sentimen berdasarkan aspek tertentu yang dibahas dalam suatu teks. Berbeda dengan analisis sentimen konvensional yang hanya mengklasifikasikan polaritas sentimen secara keseluruhan, ABSA memungkinkan analisis dilakukan secara lebih rinci dengan mengaitkan sentimen terhadap aspek atau fitur tertentu dari suatu produk atau layanan. Pendekatan ini memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang mempengaruhi persepsi pengguna terhadap suatu sistem atau aplikasi[10].

Dalam pendekatan Aspect-Based Sentiment Analysis (ABSA), terdapat beberapa komponen utama yang umumnya digunakan, yaitu Aspect Category Detection (ACD) dan Aspect Sentiment Classification (ASC). ACD digunakan untuk mengidentifikasi kategori aspek yang dibahas dalam teks ulasan pengguna, seperti performa aplikasi, kualitas jawaban, kemudahan penggunaan, atau stabilitas

sistem. ASC digunakan untuk menentukan polaritas sentimen terhadap aspek yang telah diidentifikasi, apakah bersifat positif, negatif, atau netral[38].

Namun demikian, tidak semua penelitian ABSA mengimplementasikan seluruh komponen tersebut secara lengkap. Dalam penelitian ini, pendekatan yang digunakan adalah category-based ABSA yang berfokus pada dua komponen utama, yaitu ACD dan ASC. Proses ACD dilakukan menggunakan metode Latent Dirichlet Allocation (LDA) untuk mengidentifikasi aspek secara otomatis dari data ulasan pengguna, sedangkan proses ASC dilakukan menggunakan model klasifikasi berbasis Transformer untuk menentukan sentimen pada setiap aspek yang teridentifikasi[39].

Melalui pendekatan tersebut, ABSA mampu memberikan informasi yang lebih rinci dibandingkan analisis sentimen konvensional karena tidak hanya menunjukkan apakah suatu ulasan bersifat positif atau negatif, tetapi juga menjelaskan aspek spesifik yang menjadi sumber sentimen tersebut. Dalam konteks penelitian ini, ABSA digunakan untuk mengidentifikasi berbagai aspek yang muncul dalam ulasan pengguna aplikasi ChatGPT serta menentukan polaritas sentimen terhadap setiap aspek tersebut. Dengan demikian, analisis yang dihasilkan dapat memberikan gambaran yang lebih komprehensif mengenai pengalaman dan persepsi pengguna terhadap fitur-fitur dalam aplikasi ChatGPT.

2.2.5 Inter-Rater Agreement

Inter-rater agreement merupakan metode statistik yang digunakan untuk mengukur tingkat kesepakatan antar penilai atau sistem dalam memberikan label atau keputusan terhadap data yang sama[40]. Dalam konteks pelabelan data berbasis Large Language Model (LLM), *inter-rater agreement* digunakan untuk mengevaluasi konsistensi label yang dihasilkan oleh beberapa model LLM yang berbeda tanpa memerlukan anotasi manual sebagai acuan.

Pendekatan ini menjadi penting dalam penelitian berbasis *automatic labeling*, karena kualitas label tidak dapat divalidasi secara langsung oleh annotator manusia. Oleh karena itu, tingkat kesepakatan antar model digunakan sebagai indikator reliabilitas label yang dihasilkan[41].

Beberapa metrik yang umum digunakan untuk mengukur *inter-rater agreement* antara lain sebagai berikut:

1. **Percentage Agreement**

   Percentage Agreement merupakan metrik paling sederhana yang mengukur proporsi kesamaan label antar penilai. Metrik ini dihitung sebagai persentase jumlah label yang identik dibandingkan dengan total data. Meskipun mudah diinterpretasikan, metrik ini memiliki keterbatasan karena tidak mempertimbangkan kemungkinan kesepakatan yang terjadi secara kebetulan[42].

2. **Cohen's Kappa**

   Cohen's Kappa merupakan metrik yang mengukur tingkat kesepakatan antara dua penilai dengan mempertimbangkan kemungkinan kesepakatan yang terjadi secara kebetulan[43]. Nilai Kappa dihitung menggunakan persamaan berikut:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Rumus 2.1 *Cohen's Kappa*

di mana $P_o$adalah *observed agreement* (tingkat kesepakatan aktual), dan $P_e$adalah *expected agreement* (tingkat kesepakatan yang diharapkan secara acak). Nilai Cohen's Kappa berkisar antara -1 hingga 1, dengan interpretasi sebagai berikut:

- < 0.2 : sangat lemah
- 0.2 – 0.4 : lemah
- 0.4 – 0.6 : moderat
- 0.6 – 0.8 : kuat
- 0.8 : sangat kuat

3. **Fleiss' Kappa**

   Fleiss' Kappa merupakan pengembangan dari Cohen's Kappa yang digunakan untuk mengukur kesepakatan antara lebih dari dua penilai secara simultan. Metrik ini menghitung tingkat kesepakatan keseluruhan dengan mempertimbangkan distribusi label dan kemungkinan kesepakatan acak.

Dalam penelitian ini, ketiga metrik tersebut digunakan untuk mengevaluasi konsistensi label sentimen yang dihasilkan oleh beberapa model LLM dalam *multi-LLM consistency study*. Dengan demikian, reliabilitas label yang digunakan sebagai data pelatihan dapat divalidasi secara lebih objektif tanpa bergantung pada anotasi manual.

2.2.6 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) merupakan salah satu metode *topic modeling* yang digunakan untuk mengidentifikasi topik atau pola tersembunyi dalam kumpulan dokumen teks. LDA bekerja dengan mengasumsikan bahwa setiap dokumen terdiri dari beberapa topik, dan setiap topik direpresentasikan sebagai distribusi probabilitas dari kata-kata tertentu. Dengan pendekatan ini, kata-kata yang sering muncul secara bersamaan akan dikelompokkan ke dalam satu topik yang sama[44].

Dalam konteks analisis sentimen berbasis aspek, LDA digunakan untuk melakukan penemuan aspek secara semi-otomatis (*semi-automated aspect discovery*) dari data teks tanpa memerlukan data berlabel awal. Setiap topik yang dihasilkan oleh model LDA diinterpretasikan sebagai aspek yang relevan dalam ulasan pengguna, sehingga proses identifikasi aspek dapat dilakukan secara efisien pada dataset dalam jumlah besar[31].

Topik-topik yang dihasilkan kemudian dianalisis dan diinterpretasikan secara manual untuk menentukan label aspek yang sesuai. Dengan demikian, LDA berperan sebagai metode utama dalam proses ekstraksi aspek sebelum dilakukan analisis sentimen pada tahap selanjutnya.

2.2.7 *Large Language Model* (LLM)

Large Language Model (LLM) merupakan jenis model bahasa dalam bidang Natural Language Processing (NLP) yang dikembangkan menggunakan arsitektur deep learning, khususnya berbasis Transformer, dan dilatih pada dataset teks dalam skala sangat besar. LLM dirancang untuk memahami, menghasilkan, serta memproses bahasa manusia secara kontekstual, sehingga mampu menangani

berbagai tugas NLP seperti klasifikasi teks, penerjemahan, peringkasan, hingga analisis sentimen[45].

Salah satu keunggulan utama LLM adalah kemampuannya dalam memahami konteks bahasa secara lebih mendalam dibandingkan metode tradisional. Hal ini dimungkinkan melalui mekanisme self-attention yang memungkinkan model untuk menangkap hubungan antar kata dalam suatu kalimat secara simultan. Dengan kemampuan tersebut, LLM dapat digunakan untuk menghasilkan prediksi atau label yang lebih fleksibel dan adaptif terhadap berbagai jenis data teks[46].

Penggunaan LLM dalam proses pelabelan juga mendukung pendekatan supervised learning pada tahap selanjutnya, di mana hasil labeling yang dihasilkan digunakan sebagai data latih untuk model klasifikasi berbasis Transformer. Dengan demikian, LLM berperan sebagai komponen penting dalam pipeline penelitian untuk menghasilkan dataset berlabel yang berkualitas serta mendukung proses analisis sentimen secara lebih efektif.

2.2.8 Model Transformer: IndoBERT dan XLM-RoBERTa

IndoBERT dan XLM-RoBERTa merupakan model bahasa berbasis arsitektur Transformer yang banyak digunakan dalam berbagai tugas Natural Language Processing. IndoBERT merupakan model bahasa yang dikembangkan khusus untuk bahasa Indonesia dan dilatih menggunakan korpus teks berbahasa Indonesia sehingga memiliki kemampuan yang baik dalam memahami konteks bahasa Indonesia[47].

Sementara itu, XLM-RoBERTa merupakan model multilingual yang dilatih menggunakan dataset multibahasa berskala besar. Model ini memiliki kemampuan untuk memahami berbagai bahasa sekaligus, sehingga dapat digunakan untuk berbagai tugas NLP termasuk analisis sentimen[25]. Dalam penelitian ini, kedua model tersebut digunakan untuk melakukan klasifikasi sentimen pada ulasan pengguna aplikasi ChatGPT, sehingga memungkinkan dilakukan perbandingan performa antara model monolingual dan model multilingual.

2.2.9 Metrik Evaluasi Model

Metrik evaluasi model digunakan untuk mengukur performa model dalam melakukan klasifikasi sentimen pada data teks. Evaluasi ini bertujuan untuk mengetahui seberapa baik model dalam memprediksi label sentimen pada data yang belum pernah digunakan dalam proses pelatihan[48]. Dalam penelitian klasifikasi teks, evaluasi model umumnya dilakukan dengan menggunakan beberapa metrik evaluasi yang dapat menggambarkan performa model secara menyeluruh[49].

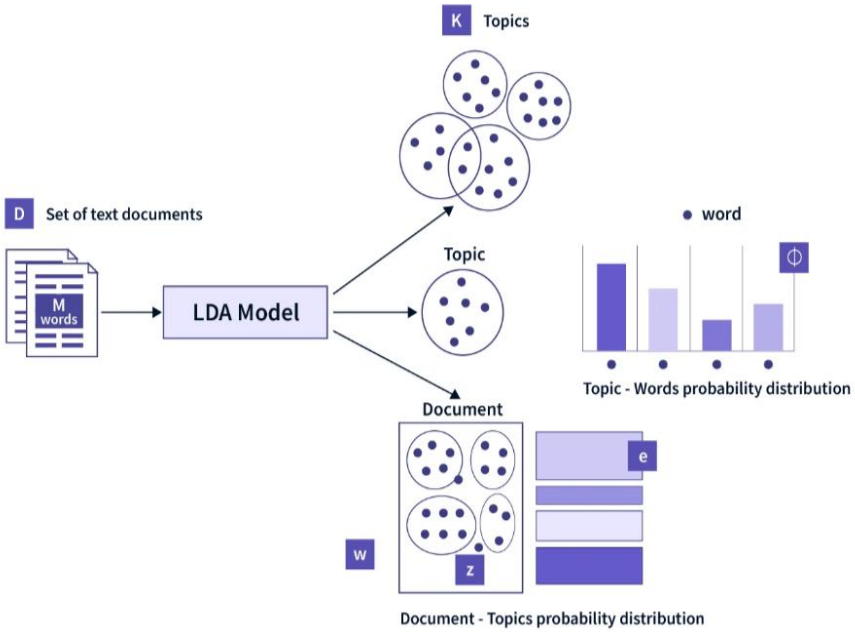
Beberapa metrik evaluasi yang umum digunakan dalam tugas klasifikasi sentimen antara lain accuracy, precision, recall, dan F1-score. Accuracy digunakan untuk mengukur proporsi jumlah prediksi yang benar terhadap seluruh data yang diuji. Precision menunjukkan tingkat ketepatan model dalam memprediksi suatu kelas tertentu, yaitu rasio antara jumlah prediksi positif yang benar dengan seluruh prediksi positif yang dihasilkan oleh model. Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang benar-benar termasuk dalam kelas tertentu. Sementara itu, F1-score merupakan rata-rata harmonis antara precision dan recall yang digunakan untuk menilai keseimbangan performa model, terutama pada dataset yang memiliki distribusi kelas yang tidak seimbang[50].

2.3 Framework/Algoritma yang digunakan

Dalam penelitian ini, pendekatan Aspect-Based Sentiment Analysis (ABSA) digunakan sebagai kerangka analisis untuk mengidentifikasi aspek dan menentukan sentimen terhadap ulasan pengguna. Implementasi dilakukan melalui ekstraksi aspek menggunakan LDA, pelabelan sentimen menggunakan LLM, serta klasifikasi menggunakan model Transformer.

Oleh karena itu, teori ABSA ini diadopsi dalam penelitian sebagai dasar keputusan metodologis untuk membedah ulasan aplikasi ChatGPT ke dalam aspek yang lebih mendetail (seperti performa sistem, kualitas konten, atau antarmuka). Langkah ini diambil sebagai argumen utama bahwa ulasan pengguna tidak dapat dinilai secara hitam-putih, melainkan harus dipetakan secara mendalam guna menghasilkan analisis preferensi yang akurat dan objektif.

2.3.1 Latent Dirichlet Allocation (LDA)



Gambar 2.1 Latent Dirichlet Allocation [44]

Latent Dirichlet Allocation (LDA) digunakan dalam penelitian ini sebagai metode untuk melakukan identifikasi aspek secara otomatis dari data ulasan pengguna dalam kerangka *Aspect-Based Sentiment Analysis* (ABSA). LDA merupakan teknik topic modeling berbasis probabilistik yang bekerja dengan mengasumsikan bahwa setiap dokumen terdiri dari campuran beberapa topik, dan setiap topik direpresentasikan sebagai distribusi probabilitas kata[18].

Dalam implementasinya, proses LDA diawali dengan tahap preprocessing data teks yang meliputi pembersihan teks, normalisasi, dan tokenisasi untuk menghasilkan representasi kata yang lebih terstruktur. Selanjutnya, data teks diubah ke dalam bentuk representasi numerik menggunakan pendekatan *bag-of-words*, sehingga setiap dokumen direpresentasikan sebagai vektor frekuensi kata dalam bentuk document-term matrix[8].

Model LDA kemudian diterapkan dengan menentukan jumlah topik (*number of topics*) sebagai parameter utama yang merepresentasikan jumlah aspek yang ingin diidentifikasi dalam penelitian. Proses pelatihan model dilakukan untuk mempelajari distribusi probabilitas kata pada setiap topik serta distribusi topik pada setiap dokumen. Hasil dari model ini berupa dua komponen utama, yaitu distribusi

kata pada topik (*topic-word distribution*) dan distribusi topik pada dokumen (*document-topic distribution*)[38].

Secara matematis, interaksi probabilistik generatif yang melandasi pembentukan kedua komponen distribusi tersebut dalam pemodelan aspek penelitian ini dirumuskan melalui fungsi peluang bersama (*joint distribution*) sebagai berikut:

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi | \alpha, \beta) = \prod_{j=1}^{M} p\left(\theta_j | \alpha\right) \prod_{i=1}^{N_j} p\left(z_{ji} | \theta_j\right) p\left(w_{ji} | \phi_{z_{ji}}\right) \prod_{k=1}^{K} p\left(\phi_k | \beta\right)$$

Rumus 2.1 Latent Dirichlet Allocation (LDA) [18]

Di mana variabel-variabel di atas didefinisikan sebagai berikut:

1. $\mathbf{w}$: Vektor kata-kata yang teramati (*observed words*) dari seluruh dokumen ulasan.
2. $\mathbf{z}$: Variabel laten yang merepresentasikan alokasi topik/aspek untuk setiap kata.
3. θ: Distribusi topik untuk setiap dokumen (*document-topic distribution*).
4. $\emptyset$: Distribusi kata untuk setiap topik (*topic-word distribution*).
5. M: Total jumlah dokumen ulasan pengguna ChatGPT yang dianalisis dalam korpus.
6. N_j: Jumlah kata di dalam dokumen ulasan ke$-j$.
7. K: Jumlah total aspek yang ditentukan secara optimal (dalam penelitian ini dievaluasi sebanyak 6 aspek).
8. $\propto$: Parameter *Dirichlet prior* yang mengontrol kerapatan campuran topik dalam suatu ulasan.
9. β: Parameter *Dirichlet prior* yang mengontrol kerapatan komponen kata dalam suatu aspek.

Berdasarkan distribusi tersebut, setiap dokumen ulasan memiliki probabilitas terhadap seluruh topik yang ada. Topik dengan nilai probabilitas tertinggi dipilih sebagai topik dominan (*dominant topic*) yang merepresentasikan

isi utama dari dokumen tersebut. Topik dominan inilah yang kemudian digunakan sebagai dasar dalam penentuan aspek pada setiap ulasan.

Untuk meningkatkan interpretasi hasil, setiap topik dianalisis berdasarkan kata-kata dengan bobot tertinggi, kemudian diberikan label aspek secara manual sesuai dengan konteks kata yang muncul, seperti performa aplikasi, kualitas jawaban, atau fitur aplikasi. Proses ini bersifat interpretatif, namun tetap berbasis pada hasil distribusi probabilitas yang dihasilkan oleh model.

Pemilihan LDA dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, LDA merupakan metode *unsupervised learning* yang tidak memerlukan data berlabel, sehingga sesuai untuk dataset ulasan pengguna yang tidak memiliki anotasi aspek. Kedua, LDA mampu menangkap pola kemunculan kata dalam skala besar, sehingga efektif digunakan pada dataset dengan jumlah data yang besar. Ketiga, pendekatan ini memungkinkan identifikasi aspek dilakukan secara data-driven, sehingga aspek yang dihasilkan lebih mencerminkan persepsi pengguna secara aktual dibandingkan pendekatan manual.

Dengan demikian, LDA berperan sebagai komponen utama dalam proses *Aspect Category Detection* (ACD) yang menghasilkan aspek-aspek secara otomatis, yang selanjutnya digunakan sebagai dasar dalam proses analisis sentimen pada tahap berikutnya. Karakteristik ekstraksi otomatis tanpa panduan (*unsupervised*) pada algoritma LDA inilah yang melandasi keputusan peneliti untuk menggunakannya dalam tahapan penentuan aspek. Teori distribusi probabilitas kata pada LDA dimanfaatkan untuk mengekstraksi topik langsung dari ribuan data ulasan mentah, sehingga menghasilkan label aspek definitif yang objektif sekaligus mengatasi kelemahan penentuan aspek secara manual yang rawan bias pada penelitian terdahulu.

2.3.2 *Large Language Model* (LLM)

Large Language Model (LLM) digunakan dalam penelitian ini sebagai pendekatan untuk melakukan pelabelan sentimen secara otomatis terhadap data ulasan pengguna aplikasi ChatGPT. Model yang digunakan adalah Qwen3-32B,

yaitu model bahasa berbasis arsitektur Transformer yang memiliki kemampuan dalam memahami konteks bahasa secara mendalam[46].

Dalam implementasinya, Qwen3-32B melalui Groq API digunakan untuk menghasilkan label sentimen pada setiap ulasan yang telah melalui tahap preprocessing. Teks ulasan diberikan sebagai input ke dalam model dalam bentuk prompt, kemudian model menghasilkan output berupa kategori sentimen, yaitu positif, negatif, atau netral. Proses ini memanfaatkan kemampuan model dalam memahami struktur kalimat serta pola bahasa seperti negasi, kontras, dan intensitas opini yang sering muncul dalam ulasan pengguna.

Tahapan penggunaan LLM dalam penelitian ini dimulai dari pemberian input berupa teks ulasan, kemudian model memproses teks tersebut untuk menghasilkan prediksi sentimen. Hasil prediksi ini selanjutnya disimpan sebagai label sentimen pada dataset. Label yang dihasilkan oleh LLM kemudian digunakan sebagai data berlabel dalam proses supervised learning pada tahap klasifikasi sentimen menggunakan model berbasis Transformer.

Penggunaan Qwen3-32B melalui Groq API dalam penelitian ini dipilih karena kemampuannya dalam menghasilkan label sentimen secara otomatis dengan pemahaman konteks yang baik, sehingga dapat menggantikan proses anotasi manual yang membutuhkan waktu dan sumber daya yang besar. Namun demikian, penggunaan LLM juga memiliki keterbatasan, seperti batasan jumlah token dalam satu kali pemrosesan serta kebutuhan komputasi yang relatif tinggi[12].

Untuk mengatasi keterbatasan tersebut, proses pelabelan menggunakan LLM dilakukan pada sebagian data sebagai data awal, kemudian hasil pelabelan tersebut digunakan untuk melatih model klasifikasi berbasis supervised learning agar dapat menangani dataset dalam skala yang lebih besar secara lebih efisien. Dengan demikian, LLM berperan sebagai sumber anotasi awal dalam pipeline penelitian yang menghubungkan tahap ekstraksi aspek dengan tahap klasifikasi sentimen.

Pendekatan ini memungkinkan integrasi yang kuat antara metode *unsupervised learning* dan *supervised learning*, di mana LDA digunakan untuk

mengidentifikasi kelompok aspek secara semi-otomatis, sedangkan LLM digunakan untuk menghasilkan label sentimen yang selanjutnya digunakan dalam proses pelatihan model klasifikasi. Dalam metodologi penelitian ini, kemampuan penalaran kontekstual mendalam dari model Qwen diintegrasikan secara terarah sebagai strategi *automated pseudo-labeling*. Teori LLM ini diterapkan untuk mengotomatisasi proses pelabelan sentimen pada dataset ulasan berskala besar, yang menjadi solusi ilmiah untuk memangkas keterbatasan waktu dan inkonsistensi jika proses anotasi dilakukan secara manual oleh manusia (*human annotator*).

2.3.3 Model Transformer

Model *Transformer* merupakan arsitektur dasar yang digunakan pada IndoBERT dan XLM-RoBERTa. Arsitektur ini menggunakan mekanisme *self-attention* untuk mempelajari hubungan antar token dalam suatu urutan teks. Melalui mekanisme tersebut, setiap token dapat memperhatikan token lain dalam kalimat, sehingga model mampu menghasilkan representasi bahasa yang lebih kontekstual.

Secara matematis, mekanisme *scaled dot-product attention* dirumuskan sebagai berikut:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Rumus 2.2 Scaled Dot-Product Attention [51]

Di mana variabel-variabel tersebut didefinisikan sebagai berikut:

- Q: Matriks *Query* yang mewakili token atau kata yang saat ini sedang dievaluasi.
- K: Matriks *Key* yang menyimpan informasi seluruh token dalam kalimat untuk dicocokkan dengan *Query*.
- V: Matriks *Value* yang berisi nilai informasi semantik aktual dari setiap token.
- d_k: Dimensi dari vektor *Key* (skala pembagi $\sqrt{d_k}$ berfungsi untuk mengontrol nilai *dot-product* agar tidak terlalu besar demi menghindari masalah gradien mengecil atau *vanishing gradient*).

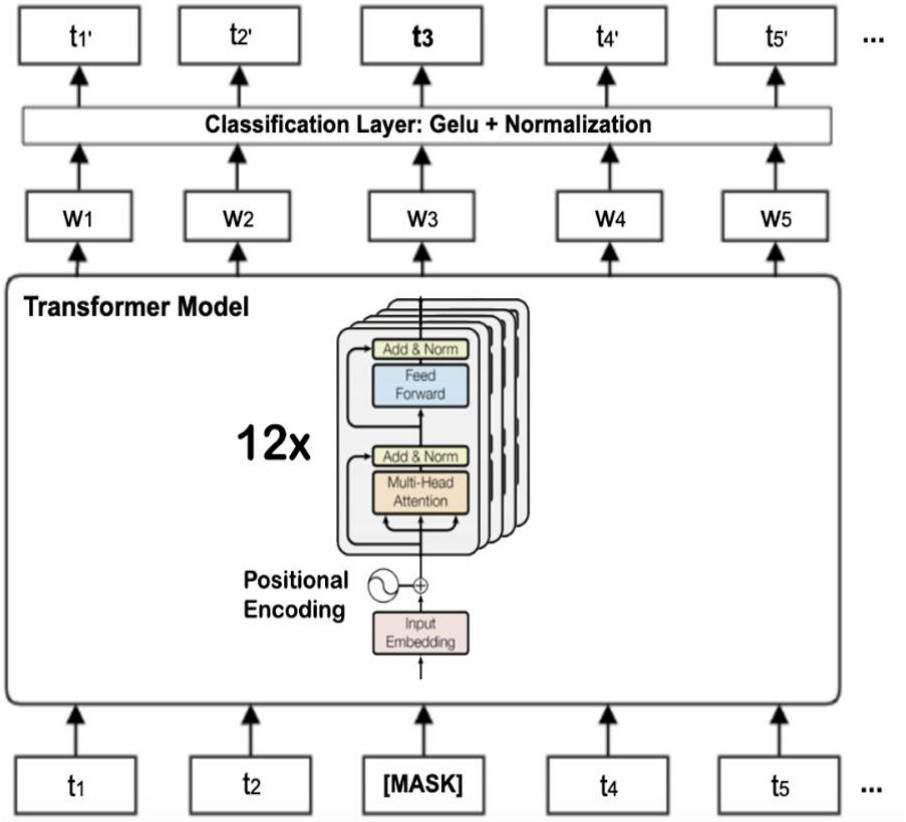
Rumus tersebut digunakan sebagai dasar mekanisme *attention* pada model berbasis *Transformer*, termasuk IndoBERT dan XLM-RoBERTa. Dengan demikian, perbedaan antara IndoBERT dan XLM-RoBERTa tidak terletak pada rumus *attention*-nya, melainkan pada cakupan bahasa, korpus pelatihan, strategi *pretraining*, serta tokenizer yang digunakan.

2.3.3.1 IndoBERT

IndoBERT merupakan *pretrained language model* untuk Bahasa Indonesia yang dikembangkan berdasarkan arsitektur *Bidirectional Encoder Representations from Transformers* (BERT). Model ini dirancang untuk menghasilkan representasi bahasa yang kontekstual dengan mempertimbangkan hubungan kata dalam dua arah (*bidirectional context*)[24]. Dengan pendekatan ini, makna suatu kata tidak dipahami secara terpisah, melainkan ditentukan oleh konteks kata sebelum dan sesudahnya dalam satu kalimat.

Pendekatan kontekstual ini sangat penting dalam analisis teks, khususnya pada ulasan pengguna aplikasi. Dalam ulasan aplikasi, sebuah kata dapat memiliki makna yang berbeda bergantung pada konteks kalimatnya. Sebagai contoh, kata "bagus" dapat menunjukkan sentimen positif dalam kalimat "aplikasi ini bagus", namun dapat berubah menjadi bernuansa negatif dalam kalimat "bagus tapi sering error". Oleh karena itu, model bahasa yang mampu memahami konteks kalimat secara menyeluruh diperlukan untuk menghasilkan interpretasi makna yang lebih[52].

Secara arsitektural, IndoBERT menggunakan konfigurasi *BERT-base* yang terdiri dari 12 lapisan Transformer encoder, 12 attention heads pada setiap lapisan, serta dimensi hidden layer sebesar 768. Setiap lapisan Transformer mengandung mekanisme *multi-head self-attention* yang berfungsi untuk mempelajari hubungan antar token dalam suatu urutan teks[9]. Melalui mekanisme ini, setiap kata dalam kalimat dapat memperhatikan kata lainnya sehingga model dapat menangkap hubungan semantik yang kompleks.



Gambar 2.2 Arsitektur BERT-base berbasis Transformer [53]

Setiap lapisan *Transformer encoder* pada IndoBERT menggunakan mekanisme *self-attention* sebagaimana dijelaskan pada Rumus 2.2. Mekanisme tersebut memungkinkan model mempelajari hubungan antar token dalam teks dan menghasilkan representasi kontekstual yang digunakan pada proses klasifikasi. Oleh karena itu, bagian ini tidak menuliskan ulang rumus *attention*, tetapi berfokus pada karakteristik IndoBERT sebagai model bahasa monolingual untuk Bahasa Indonesia.

Persamaan tersebut menunjukkan bahwa bobot perhatian dihitung melalui operasi perkalian antara matriks query dan key, kemudian dinormalisasi menggunakan fungsi softmax. Hasil normalisasi tersebut selanjutnya dikalikan dengan matriks value untuk menghasilkan representasi kontekstual dari setiap token.

Berbeda dengan model berbasis Recurrent Neural Network (RNN) yang memproses teks secara berurutan, arsitektur Transformer pada IndoBERT

memproses seluruh urutan token secara paralel sehingga memungkinkan pemahaman konteks dua arah secara lebih efektif [55]. Panjang maksimum urutan teks yang dapat diproses oleh model ini adalah 512 token.

Dalam proses representasi kalimat, IndoBERT menggunakan token khusus [CLS] yang ditempatkan pada awal urutan teks. Token ini berfungsi sebagai representasi keseluruhan kalimat (sentence representation). Representasi akhir token tersebut kemudian digunakan sebagai masukan pada lapisan klasifikasi untuk berbagai tugas pemrosesan bahasa alami seperti klasifikasi teks dan analisis sentimen.

Pada tahap pretraining, IndoBERT dilatih menggunakan pendekatan self-supervised learning dengan dua tugas utama yaitu Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP). Setelah proses pretraining selesai, model dapat digunakan pada berbagai tugas NLP melalui proses fine-tuning dengan menambahkan lapisan klasifikasi pada bagian akhir model.

Dalam penelitian ini, IndoBERT digunakan sebagai model klasifikasi sentimen untuk menganalisis ulasan pengguna berbahasa Indonesia, dengan tujuan mengevaluasi performa model monolingual dalam memahami konteks bahasa secara spesifik.

2.3.3.2 XLM-RoBERTa

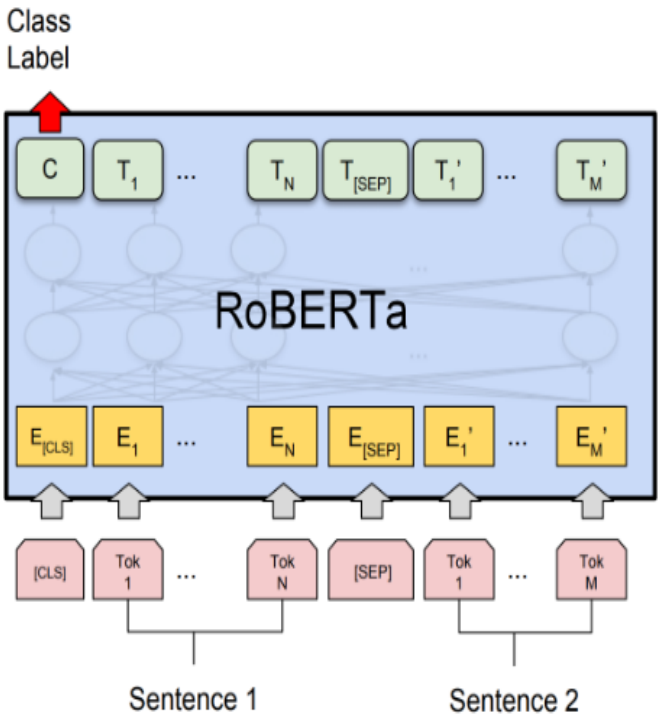
XLM-RoBERTa (Cross-lingual Language Model RoBERTa) merupakan pretrained language model berbasis arsitektur Transformer yang dikembangkan untuk mendukung pemrosesan bahasa secara multilingual. Model ini merupakan pengembangan dari arsitektur RoBERTa (Robustly Optimized BERT Approach) yang dilatih menggunakan dataset berskala besar yang mencakup lebih dari 100 bahasa[56].

Berbeda dengan BERT yang menggunakan dua tugas pretraining yaitu Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP), arsitektur RoBERTa menghilangkan tugas NSP dan berfokus pada optimalisasi tugas MLM[21]. Pendekatan ini bertujuan untuk meningkatkan kualitas representasi bahasa yang dihasilkan oleh model.

Selain itu, RoBERTa juga memperkenalkan teknik dynamic masking, yaitu strategi masking token yang berubah pada setiap epoch pelatihan. Berbeda dengan static masking pada BERT, teknik ini memberikan variasi konteks yang lebih luas sehingga model dapat mempelajari representasi bahasa secara lebih efektif.

Secara arsitektural, XLM-RoBERTa menggunakan konfigurasi yang serupa dengan BERT-base, yaitu terdiri dari 12 lapisan Transformer encoder, 12 attention heads, serta hidden dimension sebesar 768[57]. Model ini memanfaatkan mekanisme multi-head self-attention untuk memahami hubungan kontekstual antar token dalam suatu kalimat.

XLM-RoBERTa juga menggunakan mekanisme *self-attention* pada lapisan *Transformer encoder* sebagaimana dijelaskan pada Rumus 2.2. Dengan demikian, rumus dasar *attention* yang digunakan sama dengan IndoBERT. Perbedaan utama XLM-RoBERTa terletak pada cakupan bahasa, strategi *pretraining*, korpus pelatihan, dan tokenizer yang digunakan.



Gambar 2.3 Arsitektur RoBERTa berbasis Transformer [33]

Berdasarkan Gambar 2.3, model RoBERTa memproses urutan token yang telah direpresentasikan dalam bentuk embedding, kemudian diproses melalui beberapa lapisan transformer encoder untuk menghasilkan representasi kontekstual dari setiap token. Representasi token khusus [CLS] digunakan sebagai representasi keseluruhan kalimat yang selanjutnya diproses pada lapisan klasifikasi untuk menghasilkan prediksi label. Arsitektur ini juga digunakan pada model XLM-RoBERTa, yang merupakan pengembangan RoBERTa dengan kemampuan pemrosesan multibahasa[11].

Keunggulan utama XLM-RoBERTa terletak pada kemampuannya dalam mempelajari representasi bahasa lintas bahasa (cross-lingual representation). Hal ini memungkinkan model untuk memahami pola bahasa yang serupa pada berbagai bahasa sekaligus meningkatkan kemampuan generalisasi model pada berbagai jenis teks[58].

Dalam penelitian ini, XLM-RoBERTa digunakan sebagai model pembanding terhadap IndoBERT untuk mengevaluasi performa model multilingual dalam memahami konteks bahasa Indonesia dalam tugas klasifikasi sentimen. Melalui pemahaman mendasar mengenai mekanisme *self-attention* pada arsitektur *Transformer* ini, model IndoBERT (spesifik bahasa) dan XLM-RoBERTa (*multilingual*) dipilih sebagai arsitektur klasifikasi utama dalam skripsi ini. Keputusan komparasi kedua model dilakukan secara sengaja untuk membangun argumen kritis mengenai model mana yang lebih tangguh dan adaptif dalam mempelajari pola hubungan teks terhadap aspek dan sentimen pada karakteristik ulasan bahasa Indonesia yang kaya akan bahasa informal.

2.3.3.3 Transformer-based Sentiment Classification

Klasifikasi sentimen merupakan salah satu tugas dalam bidang Natural Language Processing (NLP) yang bertujuan untuk mengidentifikasi opini atau sikap yang terkandung dalam suatu teks. Dalam penelitian ini, klasifikasi sentimen digunakan untuk memetakan ulasan pengguna aplikasi ChatGPT ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral[39].

Pada pendekatan konvensional, klasifikasi sentimen biasanya dilakukan menggunakan fitur tekstual seperti bag-of-words, n-gram, atau TF-IDF. Pendekatan tersebut memiliki keterbatasan karena representasi kata bersifat statis dan tidak mempertimbangkan konteks kalimat secara menyeluruh.

Model berbasis Transformer mengatasi keterbatasan tersebut dengan memanfaatkan mekanisme self-attention untuk mempelajari hubungan antar kata dalam suatu kalimat. Melalui mekanisme ini, model dapat memahami konteks kalimat secara lebih komprehensif serta mengenali pola bahasa yang kompleks seperti negasi, kontras, maupun intensitas sentimen[33].

Dalam model berbasis Transformer seperti IndoBERT dan XLM-RoBERTa, proses klasifikasi sentimen dilakukan dengan memanfaatkan representasi global kalimat yang dihasilkan oleh token [CLS]. Representasi tersebut kemudian diproses oleh lapisan klasifikasi untuk menghasilkan probabilitas kelas sentimen[59].

Secara umum, alur klasifikasi sentimen berbasis Transformer dalam penelitian ini meliputi beberapa tahapan utama, yaitu:

1. Input teks ulasan
   Teks ulasan pengguna dimasukkan sebagai data masukan dalam model.
2. Tokenisasi teks
   Teks dipecah menjadi token menggunakan tokenizer yang sesuai dengan model yang digunakan.
3. Encoding token
   Token diubah menjadi representasi numerik berupa input IDs dan attention mask.
4. Pemrosesan oleh Transformer
   Token diproses melalui beberapa lapisan Transformer encoder untuk menghasilkan representasi kontekstual.
5. Representasi kalimat ([CLS])
   Representasi token [CLS] digunakan sebagai representasi keseluruhan kalimat.

6. Lapisan klasifikasi

   Representasi kalimat diproses oleh lapisan klasifikasi untuk menghasilkan prediksi kelas sentimen.

Pada lapisan klasifikasi akhir (tahap 6), vektor representasi kalimat dari token [CLS] (disebut sebagai $h_{[CLS]}$ dikalikan dengan matriks bobot (W) dan ditambah dengan bias (b) untuk menghasilkan nilai skor mentah (*logits*). Nilai *logits* ini kemudian diubah menjadi distribusi nilai probabilitas menggunakan fungsi Softmax berikut:

$$P(y = c|x) = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}$$

Rumus 2.3 Softmax [59]

Di mana:

- z_c : Nilai *logits* untuk kelas sentimen tertentu c.
- C: Total jumlah kategori kelas target sentimen (dalam penelitian ini terdapat 3 kelas: Positif, Negatif, dan Netral).
- $P(y = c|x)$: Nilai probabilitas hasil tebakan model bahwa teks ulasan x masuk ke dalam kelas sentimen c.

Pada tahap pelatihan (*training*), parameter internal model (bobot dan bias) akan dioptimalkan secara iteratif menggunakan algoritma AdamW dengan cara meminimalkan nilai kesalahan klasifikasi melalui fungsi Cross-Entropy Loss sebagai berikut:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(P(y = c|x))$$

Rumus 2.4 Cross-Entropy Loss [20]

Di mana:

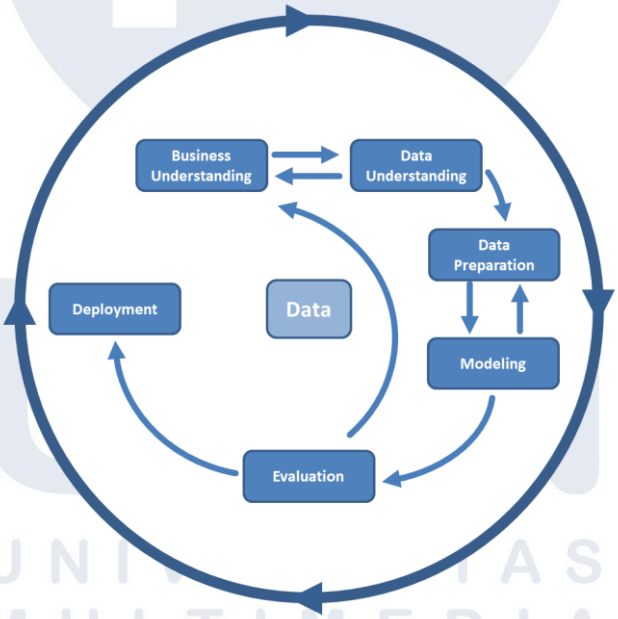
- $\mathcal{L}$: Nilai *loss* (kerugian/error) yang dihasilkan dalam satu iterasi latihan.
- y_c : Indikator biner 0 atau 1 yang menunjukkan label asli yang sebenarnya (*ground truth*). Nilainya 1 jika kelas c adalah label asli, dan 0 jika salah.

Melalui proses fine-tuning, pretrained language model dapat menyesuaikan parameter internalnya agar sesuai dengan karakteristik dataset ulasan yang

digunakan dalam penelitian. Dengan memanfaatkan representasi kontekstual yang dihasilkan oleh model Transformer, pendekatan ini mampu mengenali pola sentimen yang kompleks dalam teks ulasan pengguna secara lebih efektif dibandingkan metode berbasis fitur manual.

2.3.4 CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM (Cross-Industry Standard Process for Data Mining) merupakan kerangka kerja yang umum digunakan dalam proses penelitian dan pengembangan sistem berbasis data mining dan machine learning. Metodologi ini diperkenalkan pada tahun 1996 oleh konsorsium perusahaan yang terdiri dari SPSS, Daimler-Benz, dan NCR sebagai standar proses dalam pengolahan dan analisis data. CRISP-DM menyediakan pendekatan sistematis yang membantu peneliti dalam mengelola seluruh tahapan penelitian berbasis data secara terstruktur, mulai dari pemahaman masalah hingga implementasi hasil analisis[60].



Gambar 2.4 CRISP-DM [60]

Metodologi CRISP-DM terdiri dari enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Setiap tahapan saling berkaitan dan dapat dilakukan secara iteratif apabila diperlukan penyesuaian dalam proses penelitian.

Tahap pertama adalah Business Understanding, yaitu proses memahami permasalahan yang menjadi fokus penelitian serta menentukan tujuan yang ingin dicapai. Pada tahap ini peneliti mengidentifikasi kebutuhan analisis dan merumuskan tujuan penelitian berdasarkan konteks permasalahan yang dihadapi.

Tahap kedua adalah Data Understanding, yang bertujuan untuk memahami karakteristik dataset yang digunakan dalam penelitian. Pada tahap ini dilakukan proses pengumpulan data serta analisis eksploratif terhadap dataset untuk mengetahui struktur data, jumlah data yang tersedia, serta potensi permasalahan yang mungkin muncul pada data.

Tahap ketiga adalah Data Preparation, yaitu proses mempersiapkan dataset sebelum digunakan dalam tahap pemodelan. Pada penelitian berbasis teks, tahap ini umumnya melibatkan proses text preprocessing yang bertujuan untuk membersihkan dan menormalisasi data teks agar dapat diproses secara lebih efektif oleh model pemrosesan bahasa alami[61]. Beberapa proses yang umum dilakukan meliputi pembersihan karakter yang tidak relevan, normalisasi huruf, tokenisasi teks, serta penghapusan data duplikat. Selain itu, pada tahap ini juga dapat dilakukan proses pelabelan data serta pembagian dataset menjadi data pelatihan dan data pengujian.

Tahap berikutnya adalah Modeling, yaitu proses pengembangan model deep learning untuk menyelesaikan permasalahan yang diteliti. Pada tahap ini dataset yang telah dipersiapkan digunakan untuk melatih model agar dapat mempelajari pola yang terdapat dalam data.

Tahap kelima adalah Evaluation, yang bertujuan untuk menilai performa model yang telah dikembangkan. Evaluasi dilakukan dengan menggunakan berbagai metrik evaluasi untuk memastikan bahwa model mampu menghasilkan prediksi yang akurat dan sesuai dengan tujuan penelitian[62].

Tahap terakhir adalah Deployment, yaitu proses pemanfaatan model atau hasil analisis yang telah dikembangkan agar dapat digunakan dalam konteks yang lebih luas. Pada tahap ini, model yang telah melalui proses evaluasi dapat diimplementasikan ke dalam suatu sistem atau aplikasi sehingga dapat digunakan

secara langsung oleh pengguna. Dalam penelitian akademik, deployment dapat diwujudkan dalam berbagai bentuk, seperti penyajian hasil analisis, pengembangan aplikasi sederhana, maupun integrasi model ke dalam sistem berbasis web atau perangkat lunak tertentu untuk mendemonstrasikan kemampuan model yang telah dikembangkan.

Dalam penelitian ini, metodologi CRISP-DM digunakan sebagai kerangka kerja utama dalam proses analisis sentimen ulasan pengguna aplikasi ChatGPT yang diperoleh dari Google Play Store. Setiap tahapan CRISP-DM diterapkan secara sistematis untuk memastikan proses penelitian berjalan secara terstruktur mulai dari pengumpulan data, persiapan data, pengembangan model klasifikasi sentimen menggunakan model Transformer, hingga evaluasi performa model dan analisis hasil penelitian.

2.4 Model Evaluation

Dalam penelitian analisis sentimen berbasis model Transformer, evaluasi model merupakan tahap penting untuk menilai kinerja model dalam mengklasifikasikan sentimen ulasan pengguna. Proses evaluasi bertujuan untuk mengetahui sejauh mana model mampu melakukan prediksi secara akurat terhadap data yang belum pernah dilihat sebelumnya. Pada penelitian ini, evaluasi dilakukan terhadap dua model yang dibandingkan, yaitu IndoBERT dan XLM-RoBERTa, menggunakan dataset ulasan aplikasi ChatGPT yang diperoleh dari Google Play Store.

Proses evaluasi dilakukan menggunakan pendekatan train-test split, yaitu membagi dataset menjadi dua bagian, yaitu data pelatihan (training data) dan data pengujian (testing data). Data pelatihan digunakan dalam proses fine-tuning model, sedangkan data pengujian digunakan untuk menilai kemampuan generalisasi model terhadap data baru. Pembagian data pada penelitian ini menggunakan proporsi 80:20, di mana 80% data digunakan sebagai data pelatihan dan 20% sisanya digunakan sebagai data pengujian[63]. Pembagian data dilakukan secara acak menggunakan random seed untuk menjaga konsistensi eksperimen.

Dalam penelitian ini, klasifikasi sentimen dilakukan menggunakan tiga kelas sentimen, yaitu positif, negatif, dan netral. Oleh karena itu, metrik evaluasi dihitung menggunakan pendekatan multi-class classification, dengan menghitung nilai evaluasi pada setiap kelas sentimen.

Untuk menilai kinerja model dalam melakukan klasifikasi sentimen, digunakan beberapa metrik evaluasi yang umum digunakan dalam penelitian klasifikasi teks. Metrik evaluasi tersebut meliputi Accuracy, Precision, Recall, F1-score, serta analisis tambahan menggunakan Confusion Matrix[28].

1. Accuracy

Accuracy merupakan metrik evaluasi yang digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data pengujian. Nilai accuracy memberikan gambaran umum mengenai tingkat ketepatan model dalam melakukan klasifikasi sentimen.

Accuracy dapat dihitung menggunakan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus 2.5 Accuracy [49]

Legend dari rumus 2.2 adalah sebagai berikut:

a. **TP (True Positive)** : jumlah data positif yang diprediksi dengan benar
b. **TN (True Negative)** : jumlah data negatif yang diprediksi dengan benar
c. **FP (False Positive)** : jumlah data negatif yang keliru diprediksi sebagai positif
d. **FN (False Negative)** : jumlah data positif yang keliru diprediksi sebagai negative

2. Precision

Precision merupakan metrik evaluasi yang digunakan untuk mengukur ketepatan model dalam memprediksi suatu kelas tertentu. Precision menunjukkan seberapa banyak prediksi positif yang benar-benar merupakan data positif.

Precision dapat dihitung menggunakan persamaan berikut:

$$Precision = \frac{TP}{TP + FP}$$

Rumus 2.6 Precision [49], [62]

Legend dari rumus 2.3 adalah sebagai berikut:

a. **TP (True Positive)** : jumlah data positif yang diprediksi dengan benar
b. **FP (False Positive)** : jumlah data negatif yang keliru diprediksi sebagai positif

3. Recall

Recall merupakan metrik evaluasi yang digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang termasuk dalam kelas tertentu. Recall menunjukkan seberapa banyak data positif yang berhasil diidentifikasi oleh model dibandingkan dengan seluruh data positif yang sebenarnya.

Recall dapat dihitung menggunakan persamaan berikut:

$$Recall = \frac{TP}{TP + FN}$$

Rumus 2.7 Recall[49]

Legend dari rumus 2.4 adalah sebagai berikut:

a. TP (True Positive) : jumlah data positif yang diprediksi dengan benar
b. FN (False Negative) : jumlah data positif yang keliru diprediksi sebagai negatif

4. F1-score

F1-score merupakan metrik evaluasi yang menggabungkan precision dan recall dalam satu ukuran. Metrik ini menggunakan rata-rata harmonis antara precision dan recall sehingga dapat memberikan gambaran keseimbangan performa model.

F1-score dapat dihitung menggunakan persamaan berikut:

$$\mathrm{F1} = \frac{2 \times (Precision \times Recall)}{Precision + Recall})$$

Rumus 2.8 F1-score [49]

Legend dari rumus 2.5 adalah sebagai berikut:

a. Precision : nilai ketepatan prediksi kelas positif
b. Recall : kemampuan model dalam menemukan seluruh data positif

5. Accuracy

Selain menggunakan metrik evaluasi numerik, analisis performa model juga dapat diperkuat menggunakan confusion matrix. Confusion matrix merupakan representasi tabel yang menunjukkan jumlah prediksi benar dan salah yang dilakukan oleh model pada setiap kelas.

Confusion matrix terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Melalui confusion matrix, pola kesalahan model dapat dianalisis secara lebih mendalam, misalnya untuk mengetahui apakah ulasan negatif sering keliru diprediksi sebagai positif atau sebaliknya.

Dalam konteks penelitian ini, penggunaan confusion matrix membantu memahami perbedaan perilaku prediksi antara model IndoBERT dan XLM-RoBERTa dalam melakukan klasifikasi sentimen ulasan pengguna aplikasi ChatGPT.

2.5 Tools/software yang digunakan

2.5.1 Python

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam pengembangan aplikasi modern, khususnya pada bidang data science, machine learning, dan Natural Language Processing (NLP). Bahasa pemrograman ini dikenal memiliki sintaks yang sederhana dan mudah dipahami sehingga memungkinkan pengembang untuk membangun aplikasi dan prototipe sistem dengan lebih cepat dan efisien. Python juga mendukung berbagai paradigma pemrograman seperti pemrograman berorientasi objek, prosedural, dan fungsional.

Gambar 2.5 Python [61]

Selain itu, Python memiliki ekosistem pustaka (library) yang sangat luas yang mendukung berbagai kebutuhan komputasi, mulai dari pengolahan data hingga implementasi algoritma pembelajaran mesin. Dalam penelitian ini, Python digunakan sebagai bahasa pemrograman utama untuk mengimplementasikan seluruh proses analisis data, mulai dari pengumpulan data ulasan, pra-pemrosesan teks, pelatihan model Transformer, hingga evaluasi hasil klasifikasi sentimen.

Beberapa pustaka Python yang digunakan dalam penelitian ini antara lain Pandas untuk pengolahan dan manipulasi data, NumPy untuk komputasi numerik, Scikit-learn untuk pembagian dataset serta perhitungan metrik evaluasi, Hugging Face Transformers untuk implementasi model bahasa seperti IndoBERT dan XLM-RoBERTa, PyTorch sebagai framework pembelajaran mesin untuk proses fine-tuning model Transformer[64]. Dengan dukungan berbagai pustaka tersebut, Python memungkinkan proses analisis data dan implementasi model dilakukan secara fleksibel dan efisien.

2.5.2 Google Colaboratory

Google Colaboratory (Google Colab) merupakan platform komputasi berbasis cloud yang menyediakan lingkungan pengembangan berbasis Jupyter Notebook untuk menjalankan kode Python secara langsung melalui peramban web. Platform ini dikembangkan oleh Google dan banyak digunakan dalam penelitian di bidang machine learning dan deep learning karena menyediakan fasilitas komputasi yang cukup kuat tanpa memerlukan instalasi perangkat lunak secara lokal.

Gambar 2.6 Google Colab [62]

Salah satu keunggulan utama Google Colab adalah dukungannya terhadap penggunaan Graphics Processing Unit (GPU) dan Tensor Processing Unit (TPU) yang dapat mempercepat proses pelatihan model pembelajaran mesin. Selain itu, Google Colab juga memungkinkan integrasi langsung dengan layanan penyimpanan berbasis cloud seperti Google Drive, sehingga mempermudah proses penyimpanan dataset, kode program, maupun hasil eksperimen penelitian.

Dalam penelitian ini, Google Colab digunakan sebagai lingkungan komputasi utama untuk menjalankan seluruh proses eksperimen. Proses tersebut meliputi tahap pra-pemrosesan teks ulasan pengguna, proses tokenisasi menggunakan tokenizer yang sesuai dengan model Transformer seperti IndoBERT dan XLM-RoBERTa, proses fine-tuning model Transformer untuk melakukan klasifikasi sentimen, serta analisis lanjutan seperti ekstraksi aspek menggunakan pendekatan Aspect-Based Sentiment Analysis (ABSA).

2.5.3 Google Play Scraper

Google Play Scraper merupakan pustaka Python yang digunakan untuk mengambil data ulasan pengguna dari platform Google Play Store secara otomatis. Teknik web scraping banyak digunakan dalam penelitian berbasis data ulasan karena memungkinkan pengumpulan data dalam jumlah besar secara efisien tanpa intervensi manual [6]. Data yang diperoleh umumnya meliputi teks ulasan, rating, tanggal ulasan, serta metadata lainnya yang relevan untuk analisis sentimen.

Dalam penelitian ini, Google Play Scraper digunakan untuk mengumpulkan data ulasan pengguna aplikasi ChatGPT pada Google Play Store dengan target region Indonesia. Proses pengambilan data dilakukan secara bertahap menggunakan mekanisme *batch scraping* sehingga memungkinkan pengambilan data dalam jumlah besar. Data yang diperoleh kemudian disimpan dalam format terstruktur seperti CSV dan digunakan sebagai dataset awal dalam proses analisis sentimen. Pendekatan ini umum digunakan dalam penelitian analisis ulasan aplikasi karena mampu menyediakan data yang representatif dan relevan.

2.5.4 Gensim (Latent Dirichlet Allocation)

Gensim merupakan pustaka Python yang digunakan untuk pemodelan topik (*topic modeling*) pada data teks. Salah satu metode utama yang tersedia dalam Gensim adalah *Latent Dirichlet Allocation* (LDA), yaitu model probabilistik yang digunakan untuk mengidentifikasi topik tersembunyi dalam kumpulan dokumen secara tidak terawasi (*unsupervised learning*) [65]. Metode LDA bekerja dengan mengelompokkan kata-kata yang sering muncul bersama dalam suatu dokumen ke dalam topik tertentu, sehingga mampu merepresentasikan struktur tematik dalam data teks.

Gambar 2.7 Gensim LDA [63]

Dalam penelitian ini, Gensim digunakan untuk mengimplementasikan metode LDA dalam proses ekstraksi aspek pada ulasan pengguna. Model LDA dilatih menggunakan data teks yang telah melalui tahap preprocessing untuk menghasilkan sejumlah topik yang merepresentasikan aspek-aspek utama dalam ulasan pengguna. Penentuan jumlah topik dilakukan dengan mempertimbangkan

nilai *coherence score* untuk memperoleh hasil yang optimal. Hasil dari proses ini digunakan sebagai dasar dalam pendekatan *Aspect-Based Sentiment Analysis* (ABSA), sehingga setiap ulasan dapat dianalisis berdasarkan aspek yang relevan. Penggunaan LDA dalam analisis teks telah banyak diterapkan dalam penelitian terkini karena kemampuannya dalam mengidentifikasi pola topik secara efektif.

2.5.5 Langdetect

Langdetect merupakan pustaka Python yang digunakan untuk mendeteksi bahasa dari suatu teks secara otomatis. Pustaka ini bekerja dengan memanfaatkan distribusi karakter dan pola linguistik untuk mengidentifikasi bahasa yang digunakan dalam suatu dokumen[66]. Dalam konteks analisis teks, deteksi bahasa menjadi langkah penting terutama ketika dataset bersifat multilingual.

Dalam penelitian ini, Langdetect digunakan pada tahap preprocessing untuk menyaring ulasan pengguna berdasarkan bahasa. Mengingat dataset yang diperoleh dari Google Play Store bersifat multilingual, maka hanya ulasan berbahasa Indonesia yang dipertahankan untuk memastikan kesesuaian dengan model yang digunakan, yaitu IndoBERT dan XLM-RoBERTa. Proses ini bertujuan untuk meningkatkan kualitas data serta memastikan hasil analisis sentimen lebih relevan dan akurat. Penggunaan teknik deteksi bahasa dalam preprocessing juga telah banyak digunakan dalam penelitian analisis sentimen untuk meningkatkan performa model.

2.5.6 Groq API

Large Language Model (LLM) merupakan model berbasis deep learning yang dirancang untuk memahami, menghasilkan, dan menganalisis teks dalam bahasa alami. Model ini dilatih menggunakan data dalam jumlah besar sehingga mampu menangkap konteks, pola bahasa, serta hubungan antar kata dalam suatu kalimat secara lebih mendalam dibandingkan metode tradisional[9].

Dalam penelitian ini, digunakan model Qwen3-32B dengan ukuran parameter 32 billion parameters sebagai bagian dari proses pelabelan sentimen pada data ulasan pengguna. Model ini dimanfaatkan untuk mengidentifikasi polaritas

sentimen, yaitu positif, negatif, atau netral, berdasarkan konteks kalimat yang terdapat dalam ulasan. Dengan kemampuan pemahaman bahasa yang dimiliki, model ini mampu menangani variasi bahasa tidak formal, kesalahan penulisan, serta ekspresi opini yang kompleks.

Gambar 2.8 Groq API [64]

Proses penggunaan model Qwen3-32B dilakukan melalui layanan Groq API sebagai media akses terhadap model. Penggunaan layanan ini memungkinkan proses inferensi dilakukan secara efisien tanpa memerlukan sumber daya komputasi lokal yang besar. Dengan demikian, proses pelabelan sentimen dapat dilakukan secara otomatis dan scalable, sehingga mendukung pengolahan dataset dalam jumlah besar.

2.5.7 Hardware yang Digunakan

Penelitian ini dilakukan dengan ditenagai oleh beberapa komponen hardware, yaitu sebagai berikut:

1. Processor (CPU) : AMD Ryzen™ 5 7000 Series
2. Memory (RAM) : 16 GB
3. Operating System : Windows 11
4. Storage : 512 GB SSD