BAB 2
LANDASAN TEORI

Pada bagian ini dijabarkan teori-teori yang mendasari penelitian secara lengkap dan menyeluruh.

2.1 Penelitian Pendahulu

Sebelum melakukan penelitian ini, dilakukan telaah terhadap beberapa studi terdahulu yang berkaitan dengan analisis sentimen. Berbagai penelitian telah memanfaatkan algoritma *Bidirectional Encoder Representations from Transformers* (BERT) dan variannya seperti IndoBERT dalam mengklasifikasikan sentimen publik terhadap suatu entitas atau strategi pemasaran. Studi-studi ini menjadi referensi penting dalam merancang metodologi penelitian ini. Tabel 2.1 menyajikan ringkasan beberapa penelitian relevan yang menggunakan pendekatan BERT atau IndoBERT dalam konteks analisis sentimen.

Tabel 2.1. Penelitian pendahulu

No	Penulis	Judul	Algoritma	Hasil
1	Manoppo et al. [11]	Analisis Sentimen Publik di Media Sosial Terhadap Kenaikan PPN 12% di Indonesia Menggunakan IndoBERT	IndoBERT	Akurasi 84,94%, presisi 85,60%, *recall* 84,94%, *F1-score* 84,37%
2	Pradana [9]	Analisis Sentimen Masyarakat Media Sosial Twitter terhadap Kinerja Penjabat Gubernur DKI Jakarta Menggunakan Model IndoBERT	IndoBERT	Akurasi, presisi, *recall*, dan *F1-score* masing-masing 90.5%, 90.6%, 90.5%, 90.49% pada data uji dan 96.63%, 96.64%, 96.63%, 96.62% pada data latih

Tabel 2.1 Tabel Penelitian Pendahulu (lanjutan)

No	Penulis	Judul	Algoritma	Hasil
3	Mario et al. [12]	Analisa Sentimen terhadap *Cosplayer* pada Platform Sosial Media Instagram Menggunakan Metode *Simple Additive Weighting*	IndoBERT	Akurasi model *indobenchmark /indobert-base-p1* 76.34%, *indolem/indobe rtweet-base-uncased* 55.82%, dan *indolem/indobe rt-base-uncased* 43.31%
4	Sjoraida et al. [13]	Analisis Sentimen Film *Dirty Vote* Menggunakan BERT (*Bidirectional Encoder Representations from Transformers*)	BERT	Akurasi sebesar 85%, presisi 86%, *recall* 84%, dan *F1-score* 85%
5	Sriyanti et al. [14]	Implementasi Model BERT pada Analisis Sentimen Pengguna Twitter terhadap Aksi Boikot Produk Israel	BERT	Akurasi, presisi, *recall*, dan *F1-score* masing-masing sebesar 85%

Penelitian oleh Manoppo et al. [11] menggunakan model IndoBERT untuk melakukan analisis sentimen publik di berbagai platform media sosial (X, Instagram, dan TikTok) mengenai kenaikan Pajak Pertambahan Nilai (PPN) 12% di Indonesia. Studi ini menggunakan 2.581 sampel data dan membagi *dataset* menjadi 80/10/10 untuk pelatihan, validasi, dan pengujian. Sebelum data memasuki pelatihan, data akan dibersihkan terlebih dahulu dengan melakukan *case folding*, menghapus URL, dan melakukan normalisasi *slang*. Model IndoBERT di *fine-tune* selama tiga *epoch* dan menunjukkan kinerja yang signifikan dengan akurasi 84,94%, presisi 85,60%, *recall* 84,94%, dan *F1-score* 84,37%. Temuan ini menegaskan efektivitas IndoBERT dalam tugas klasifikasi sentimen berbahasa

Indonesia pada data media sosial, serta menunjukkan mampu memberikan analisis yang kuat terhadap opini publik.

Penelitian yang dilakukan oleh Pradana [9] bertujuan untuk mengevaluasi opini publik terhadap kinerja Penjabat Gubernur DKI Jakarta, Heru Budi Hartono, melalui analisis sentimen di Twitter menggunakan model IndoBERT. Setelah melalui tahap *preprocessing* dan pelabelan manual, model dilatih dan dievaluasi menggunakan metrik seperti akurasi, presisi, *recall*, dan *F1-score*. Hasil evaluasi menunjukkan performa yang tinggi dengan akurasi sebesar 90,5% dan *F1-score* sebesar 90,49%. Selanjutnya, model digunakan untuk memprediksi 4.650 data tidak berlabel, yang menghasilkan 2.791 data bersentimen positif dan 1.859 data negatif. Penelitian ini menjadi acuan penting dalam penerapan IndoBERT untuk analisis sentimen media sosial karena menunjukkan kemampuan model dalam menangani data riil, termasuk pelabelan otomatis data tanpa anotasi.

Penelitian oleh Mario et al. [12] bertujuan untuk menganalisis sentimen publik terhadap *cosplayer* di platform media sosial Instagram. Peneliti mengklasifikasikan *dataset* ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif. Pada penelitian ini, data komentar dikumpulkan melalui API Instagram dan diproses melalui tahapan pembersihan serta pra-pemrosesan teks. Beberapa model IndoBERT digunakan untuk klasifikasi sentimen, di antaranya `indobenchmark/indobert-base-p1` yang mencapai akurasi tertinggi sebesar 76,34%, dibandingkan dengan model lain seperti `indobertweet-base-uncased` dan `indobert-base-uncased`. Hasil ini menunjukkan bahwa model IndoBERT `indobenchmark/indobert-base-p1` unggul dalam identifikasi dan klasifikasi data sentimen.

Penelitian Sjoraida et al. [13] bertujuan untuk menganalisis sentimen publik terhadap film dokumenter *Dirty Vote* dengan menggunakan model BERT. Data diperoleh dari media sosial Twitter, kemudian melalui proses *preprocessing* dan pelabelan secara manual menjadi tiga kelas sentimen, yaitu positif, negatif, dan netral. Model BERT digunakan karena kemampuannya dalam memahami konteks kata secara mendalam melalui pendekatan *bidirectional encoding*. Hasil pelatihan dan pengujian menunjukkan bahwa model ini berhasil mengklasifikasikan sentimen dengan baik, ditunjukkan oleh nilai akurasi dan *F1-score* yang tinggi, masing-masing 85%. Penelitian ini menjadi salah satu bukti bahwa model BERT efektif digunakan dalam tugas klasifikasi sentimen berbasis teks Bahasa Indonesia, khususnya pada isu-isu sosial dan politik yang sedang berkembang.

Penelitian oleh Sriyanti et al. [14] membahas implementasi model BERT dalam

menganalisis sentimen masyarakat Twitter terhadap aksi boikot produk Israel. Data dikumpulkan melalui proses *scraping* dari Twitter, kemudian dilakukan *preprocessing* dan pelabelan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Model BERT digunakan karena kemampuannya dalam memahami konteks kalimat secara mendalam dan akurat dalam Bahasa Indonesia. Setelah proses pelatihan, model dievaluasi menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* dengan hasil masing-masing 85%. Penelitian ini mengklasifikasikan data sentimen bahasa Indonesia kemudian diuji menggunakan *confusion matrix*. Hasilnya menunjukkan bahwa model BERT mampu melakukan klasifikasi sentimen dengan hasil yang cukup baik, menjadikannya sebagai pendekatan yang efektif dalam memahami opini publik terhadap isu-isu sosial dan geopolitik. Penelitian ini menjadi salah satu referensi yang relevan dalam penggunaan model BERT untuk klasifikasi sentimen dalam Bahasa Indonesia di media sosial. Hasil dari penelitian-penelitian pendahulu membuktikan bahwa BERT merupakan model yang andal untuk analisis sentimen berbasis teks, terutama pada isu-isu sensitif dan ramai diperbincangkan di media sosial.

2.2 Korean Wave (Hallyu)

Hallyu diartikan secara harfiah sebagai gelombang Korea atau *Korean Wave* [15]. Penyebaran gelombang Korea ini bermula dari festival Piala Dunia 2002 yang diselenggarakan pertama kali di benua Asia, yaitu di Korea Selatan dan Jepang. Diadakannya festival Piala Dunia di Korea Selatan membuat citra negara ini semakin meningkat dan mulai dipandang oleh negara-negara lainnya. *Korean Wave* adalah istilah yang merujuk pada penyebaran budaya populer Korea Selatan secara global ke berbagai negara di dunia, termasuk Indonesia [16]. Kebudayaan Korea Selatan yang memadukan kehidupan tradisional dan modern ini mulai tersebar secara global melalui media, seperti film, drama, musik, makanan, *fashion*, dan produk kecantikan.

Pada jurnal yang berjudul *Past, Present and Future of Hallyu (Korean Wave)*, terdapat tiga fase persebaran gelombang Korea atau *Hallyu* ke berbagai negara [17]. Di Indonesia, tahap *Korean Wave* 1.0 dapat dilihat dengan hadirnya serial drama Korea (K-Drama) pertama kali di salah satu stasiun televisi swasta, yaitu Trans TV. K-Drama pertama yang ditayangkan adalah drama *Mother's Sea*. Setelah itu, terdapat media televisi lain yang mulai menayangkan drama-drama Korea Selatan, seperti stasiun TV Indosiar dengan drama berjudul *Endless Love*. Pada tahun 2011,

tercatat ada sekitar 50 drama Korea yang ditayangkan pada stasiun TV swasta Indonesia.

Tahap *Korean Wave* 2.0 ditandai dengan masuknya budaya K-pop ke Indonesia melalui grup-grup seperti Super Junior, Girls' Generation, Seventeen, Enhypen, dan NCT. Budaya K-pop mendapat sambutan luas, bahkan menarik perhatian berbagai kalangan, mulai dari remaja hingga ibu rumah tangga. Beberapa idola K-pop telah menggelar konser di Indonesia, dengan penjualan tiket yang selalu terjual habis atau *sold out*. Bahkan, beberapa grup mencatat penjualan tiket habis dalam hitungan detik atau menit. Fenomena ini menunjukkan antusiasme tinggi para penggemar dalam mendukung idola kesukaannya. Pada fase *Korean Wave* 3.0, penyebaran kebudayaan Korea Selatan tidak lagi terbatas pada drama Korea dan K-pop, melainkan mencakup keseluruhan produk budaya Korea, seperti makanan, *fashion*, kosmetik, dan gaya hidup.

Sebagai bagian dari gelombang *Hallyu* 2.0 dan 3.0, NCT DREAM merupakan salah satu grup idola K-Pop yang telah meraih popularitas signifikan, terutama di pasar global termasuk Indonesia. Grup ini dikenal memiliki basis penggemar yang sangat loyal dan aktif, dengan tingkat *engagement* tinggi yang terbukti dalam berbagai kampanye. Fenomena antusiasme dan loyalitas penggemar ini menjadi faktor utama pendorong strategi pemasaran berbasis kolaborasi, dengan *brand* memanfaatkan daya tarik idola K-Pop seperti NCT DREAM untuk menjangkau audiens yang luas dan loyal. Kolaborasi semacam ini tidak hanya berfungsi sebagai sarana promosi, tetapi juga menjadi jembatan efektif antara *brand* dengan komunitas penggemar yang sangat terlibat, memicu sentimen positif dan partisipasi aktif dalam kampanye pemasaran.

2.3 Analisis Sentimen

Analisis sentimen merupakan cabang dari ilmu komputasi yang bertujuan untuk memahami dan mengklasifikasikan opini atau emosi yang terkandung dalam suatu teks. Melalui pendekatan ini, sebuah kalimat atau dokumen dianalisis untuk mengidentifikasi kecenderungan sikap penulis, apakah bersifat positif, negatif, atau netral. Hasil dari analisis ini dapat digunakan dalam sistem rekomendasi maupun visualisasi data. Teknik analisis sentimen ini banyak digunakan dalam berbagai bidang seperti pemasaran, politik, dan layanan pelanggan.

Analisis ini juga dapat menggambarkan ekspresi emosional yang muncul dalam teks, seperti kebahagiaan, kemarahan, atau kesedihan, sehingga memungkinkan

kita untuk memahami bagaimana perasaan seseorang atau reaksi publik terhadap isu, produk, merek, atau tokoh tertentu yang dibahas di internet. Analisis sentimen sering disebut juga sebagai *opinion mining*, karena memungkinkan sistem untuk menilai opini subjektif seseorang terhadap suatu topik [18]. Hal ini memberikan wawasan yang penting bagi pengambil keputusan untuk memahami persepsi publik secara lebih mendalam, terutama ketika menangani isu-isu yang banyak diperbincangkan di media sosial.

2.4 Natural Language Processing

Natural Language Processing (NLP) merupakan sebuah teknik yang mengintegrasikan bidang linguistik, ilmu komputer, dan kecerdasan buatan untuk memungkinkan komputer dapat memahami, menafsirkan, dan menghasilkan bahasa manusia [19]. Tujuan dari NLP adalah untuk menciptakan interaksi yang lebih alami antara manusia dan komputer, layaknya komunikasi antar manusia pada umumnya [20]. Beberapa contoh penggunaan teknologi ini meliputi asisten virtual seperti Siri dan Alexa, sistem pencarian informasi, analisis sentimen di media sosial, serta pengklasifikasian dokumen [21]. Teknologi NLP memungkinkan sistem untuk menyesuaikan diri dengan kebutuhan pengguna secara lebih personal dan responsif, meningkatkan layanan di berbagai bidang, dan membuat interaksi dengan teknologi lebih intuitif serta bermanfaat [22].

2.5 Dataset Labeling and Inter-Analyst Validation

Pada penelitian analisis sentimen, kualitas dan konsistensi data yang diberi label secara manual merupakan faktor penentu keberhasilan model yang dibangun. Proses pelabelan data secara manual, atau yang dikenal dengan anotasi, memerlukan pedoman yang jelas dan terstruktur untuk memastikan validitasnya. Peneliti perlu memastikan bahwa pelabelan tidak berdasarkan pada penafsiran pribadi, melainkan dilakukan secara konsisten oleh semua penilai. Hal ini sangat penting untuk menjamin keandalan data dan memastikan bahwa penelitian dapat di replikasi pada masa mendatang [23].

Untuk menilai sejauh mana dua *annotator* memiliki kesepakatan dalam memberi label, penggunaan persentase kesepakatan sederhana sering kali dianggap kurang tepat. Hal ini karena metrik tersebut tidak mempertimbangkan kemungkinan kesepakatan yang terjadi secara kebetulan (*chance agreement*). Sebagai contoh,

jika salah satu kategori sentimen jauh lebih banyak daripada yang lain, para penilai bisa saja tampak seolah-olah setuju hanya karena cenderung memilih kategori yang paling sering muncul, bukan karena benar-benar memiliki pemahaman yang sama [23]. Oleh karena itu, dibutuhkan metrik yang lebih akurat untuk mengukur tingkat kesepakatan tersebut.

Salah satu metrik yang umum digunakan untuk mengukur tingkat kesepakatan yang lebih akurat adalah koefisien Cohen's Kappa. Cohen's Kappa (κ) merupakan metrik statistik yang dirancang untuk menilai tingkat keandalan antar penilai dengan mengurangi pengaruh kesepakatan yang mungkin terjadi secara kebetulan [24]. Metrik ini digunakan untuk mengetahui seberapa besar dua *annotator* benar-benar setuju dalam mengklasifikasikan data. Nilai Cohen's Kappa dihitung dengan membandingkan antara kesepakatan yang sebenarnya terjadi (P_o) dan kesepakatan yang diperkirakan terjadi secara kebetulan (P_e), dengan rumus sebagai berikut:
Rumus 2.1 menunjukkan cara perhitungan *Cohen's Kappa*.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.1}$$

Keterangan:

1. Kesepakatan yang Diamati (P_o): Proporsi dari total item di mana kedua *annotator* memberikan label yang sama.
2. Kesepakatan yang Kebetulan (P_e): Probabilitas teoritis di mana kedua anotator akan setuju secara acak, dihitung berdasarkan distribusi label yang diberikan oleh masing-masing *annotator*.

Panduan interpretasi nilai Cohen's Kappa, yang dikembangkan oleh Landis dan Koch (1977), menyajikan skala yang kredibel untuk menilai tingkat kesepakatan antar-*annotator* [25]. Tabel 2.2 menyajikan tampilan skala interpretasi tersebut.

Tabel 2.2. Skala interpretasi nilai koefisien Cohen's Kappa

Nilai Koefisien (κ)	**Tingkat Kesepakatan**
<0.00	Buruk
0.00-0.20	Ringan (*Slight*)
0.21-0.40	Wajar (*Fair*)
0.41-0.60	Moderat
0.61-0.80	Substansial
0.81-1.00	Hampir Sempurna

Nilai koefisien ini berkisar antara -1 hingga 1. Nilai 1 menunjukkan kesepakatan sempurna, 0 menunjukkan kesepakatan yang sama dengan kebetulan, dan nilai negatif menunjukkan kesepakatan yang lebih buruk dari kebetulan. Nilai Kappa di atas 0.61 umumnya dianggap menunjukkan tingkat kesepakatan yang substansial hingga hampir sempurna, menjadikan data layak untuk digunakan dalam penelitian.

2.6 Text Preprocessing

Text Preprocessing merupakan langkah awal untuk membersihkan data dari kata atau elemen tertentu yang dapat memengaruhi hasil analisis sentimen. Proses pembersihan data ini penting untuk dilakukan, sehingga kinerja model IndoBERT dapat bekerja dengan lebih baik [26]. Berikut beberapa tahapan dalam *Text Preprocessing*.

1. *Data Cleansing*

   Data cleansing atau pembersihan data merupakan proses penghapusan elemen-elemen yang tidak relevan atau tidak memberikan kontribusi signifikan terhadap analisis sentimen, seperti URL [11]. Proses pembersihan data juga menghapus *whitespace* untuk menstandarisasi format teks dan menghilangkan spasi berlebih yang dapat mengganggu analisis [27]. Namun, elemen lain seperti emoji, tanda baca, angka, *mention*, dan *hashtag* tetap dipertahankan karena dapat memuat konteks, ekspresi emosional, maupun penanda topik yang penting bagi model untuk memahami makna sebenarnya dari komentar.

   Emoji memiliki peran dan makna penting yang tidak dapat diabaikan. Penghapusan emoji secara langsung dapat menyebabkan hilangnya informasi sentimen yang krusial [27]. Oleh karena itu, penelitian ini menggunakan

metode emoji *replacement* dengan kamus yang mengubah emoji menjadi representasi kata sesuai konteks, tanpa menghapusnya. Pendekatan ini bertujuan untuk memastikan bahwa makna dan konteks sentimen yang disampaikan melalui emoji tetap terpelihara dan dapat dipahami oleh model analisis.

2. *Case Folding*
   Case folding merupakan proses untuk mengubah format huruf dalam teks menjadi huruf kecil (*lowercase*), namun tidak mengubah makna atau struktur kalimat [12]. Tahap ini dilakukan karena penggunaan huruf besar yang tidak konsisten dalam suatu teks dapat memengaruhi performa *training* data.

3. *Tokenization*
   Tokenization merupakan proses memecah teks menjadi unit-unit yang lebih kecil (token), yang membantu model dalam memahami struktur dan makna dari suatu kalimat. Penelitian ini menggunakan *tokenizer* dari IndoBERT yang berbasis WordPiece, yaitu metode tokenisasi yang mampu mengelola kata-kata langka maupun elemen non-standar seperti emoji, *hashtag*, angka, *mention*, maupun tanda baca [28]. Oleh karena itu, elemen-elemen ini tidak dihapus dalam tahap *cleaning* karena *tokenizer* IndoBERT dapat menangani elemen-elemen tersebut secara langsung. Pendekatan ini juga memastikan bahwa makna sentimen atau topik yang terkandung dalam elemen tersebut tetap dapat diproses oleh model secara kontekstual.

4. *Normalization*
   Normalisasi adalah proses untuk mengubah kata-kata tidak baku, singkatan, atau istilah *slang* menjadi bentuk kata yang baku sesuai standar dalam Bahasa Indonesia [27]. Pada penelitian ini, proses normalisasi dilakukan dengan bantuan kamus *abbreviations* yang telah disusun secara manual. Tujuannya adalah untuk memastikan bahwa model dapat memahami makna sebenarnya dari setiap kata tanpa kesalahan interpretasi akibat variasi penulisan informal yang sering muncul di media sosial.

2.7 Machine Learning

Kecerdasan buatan atau *Artificial Intelligence* (AI) merupakan salah satu bidang ilmu komputer yang digunakan untuk mengembangkan *software* dan *hardware* yang dapat meniru cara berpikir seperti manusia [29]. *Machine Learning* (ML)

adalah cabang dari AI berupa aplikasi komputer dan algoritma matematika yang memungkinkan sistem untuk belajar dari suatu data untuk kemudian membuat prediksi atau keputusan di masa yang akan datang [30]. ML bekerja dengan mengembangkan model berdasarkan pola yang didapat dari *dataset* yang telah dikumpulkan. Proses pembelajaran ML dalam memperoleh kecerdasan dilakukan dalam dua tahap, yaitu pelatihan (*training*) dan pengujian (*testing*) [31]. Secara umum, *Machine Learning* dibagi dalam tiga kategori utama, antara lain:

1. *Supervised Learning*
   Model dilatih menggunakan data yang sudah diklasifikasi dan diberikan label untuk mengenali kelas yang tidak dikenal. Umumnya digunakan dalam klasifikasi dan regresi.

2. *Unsupervised Learning*
   Model bekerja menggunakan data yang tidak memiliki label. Umumnya digunakan dalam *clustering* atau reduksi dimensi.

3. *Reinforcement Learning*
   Model belajar melalui interaksi dengan lingkungan yang dinamis untuk mendapatkan *reward* atau *punishment*.

Penelitian ini menggunakan metode *Supervised Learning* karena dalam melakukan analisis sentimen, model yang dibuat membutuhkan *dataset* berlabel yang berisi teks-teks atau kalimat yang telah dikategorikan ke dalam sentimen tertentu seperti positif, negatif, atau netral.

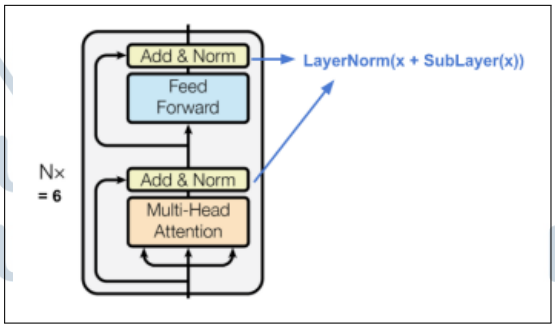
2.8 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) merupakan model berbasis *deep learning* yang dikembangkan oleh *Google* untuk meningkatkan pemahaman mengenai bahasa alami atau *Natural Language Processing* (NLP) [32]. BERT mampu membaca teks secara *bidirectional*, sehingga mampu memahami makna kata atau kalimat dalam konteks yang luas. Namun, BERT tidak dapat digunakan untuk membuat dan menerjemahkan teks karena dilatih dengan *Masked Language Model* MLM, yang memungkinkan model BERT hanya bersifat dua arah (kiri-ke-kanan dan kanan-ke-kiri) [9]. Model ini didasarkan oleh arsitektur *Transformer* dan telah digunakan pada macam-macam aplikasi NLP seperti pemodelan bahasa, menjawab pertanyaan, dan analisis sentimen. Penggunaan

algoritma BERT dalam analisis sentimen dapat memberikan pemahaman yang lebih mendalam terhadap opini dan respons pada teks, baik untuk penelitian, pengembangan produk, maupun pemahaman masyarakat terhadap suatu topik [13]. Berbagai penelitian telah menunjukkan bahwa analisis sentimen menggunakan model dari algoritma BERT mampu menunjukkan keberhasilan dan hasil akurasi yang signifikan. Selain itu, analisis sentimen menggunakan BERT juga terbukti efektif dalam mengatasi masalah seperti *slang*, kesalahan ejaan, aksen modern, dan tata bahasa [33]. Pada penelitian tentang analisis sentimen terhadap vaksinasi HPV lewat Twitter, berhasil menunjukkan bahwa penggunaan model BERT untuk analisis sentimen adalah yang paling unggul [34].

Model BERT memiliki arsitektur berlapis ganda (*multi-layer*) *bidirectional Transformers* dengan lapisan *encoder* untuk membaca suatu *input*. *Transformers* dapat mempelajari dan membentuk pemahaman konteks kata melalui mekanisme *self-attention*, yaitu teknik yang memungkinkan model memfokuskan perhatian pada kata-kata relevan dalam suatu kalimat untuk memahami makna secara kontekstual. *Transformers* memiliki dua komponen utama, namun BERT hanya menggunakan bagian *encoder*.

Encoder bertugas untuk membaca dan memahami *input* teks. Terdiri dari tumpukan N = 6 lapisan identik, yang setiap lapisannya memiliki dua sublapisan yaitu *self-attention* dan jaringan saraf *feed-forward*. Mekanisme *self-attention* memungkinkan model untuk menangkap hubungan antar kata dalam satu kalimat secara kontekstual, sehingga makna kata dapat dipahami berdasarkan posisi dan kata lain di sekitarnya, misalnya kata ”bank” dalam ”bank sungai” akan diproses berbeda dibanding ”bank uang”, karena perhatian (*attention*) diarahkan sesuai konteks. Struktur umum *encoder* pada arsitektur *Transformer* yang menjadi dasar BERT ditunjukkan pada Gambar 2.1.

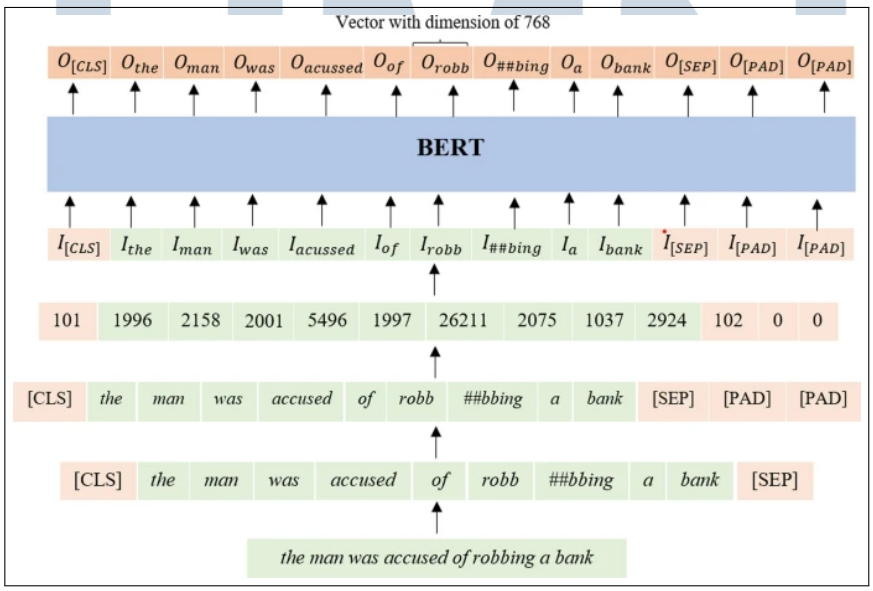


Gambar 2.1. *Encoder*

Sumber: [9]

2.8.1 Input dan Output BERT

Secara fundamental, model BERT memproses teks dengan terlebih dahulu mengubahnya menjadi representasi numerik melalui proses tokenisasi dan *encoding*. Berbeda dengan tokenisasi tradisional, BERT menggunakan metode *subword tokenization* yang memecah teks menjadi potongan-potongan kecil (*wordpiece*) [35]. Pendekatan ini memungkinkan model untuk menangani kata-kata yang belum pernah dilihat sebelumnya (*out-of-vocabulary*) secara lebih efektif.
Proses tokenisasi ini juga melibatkan penambahan token-token khusus, seperti [CLS] di awal teks (yang digunakan untuk tugas klasifikasi) dan [SEP] di akhir teks, untuk menandai batas kalimat [35]. Untuk memastikan semua teks dalam satu *batch* memiliki panjang yang seragam, digunakanlah dua mekanisme, yaitu *padding* dan *truncation*. *Padding* adalah proses menambahkan token khusus [PAD] pada akhir teks hingga panjangnya mencapai batas maksimum. Sebaliknya, *truncation* adalah proses pemotongan teks jika jumlah tokennya melebihi panjang maksimum [35]. Hal ini penting untuk menghindari *input* yang terlalu panjang sehingga tidak dapat diproses oleh model.
Tahap selanjutnya adalah *encoding*, yang bertujuan untuk memetakan token-token yang sudah disiapkan menjadi bentuk numerik agar dapat diproses oleh model BERT. Gambar 2.2 mengilustrasikan secara visual alur proses *input* dan *output* model BERT.



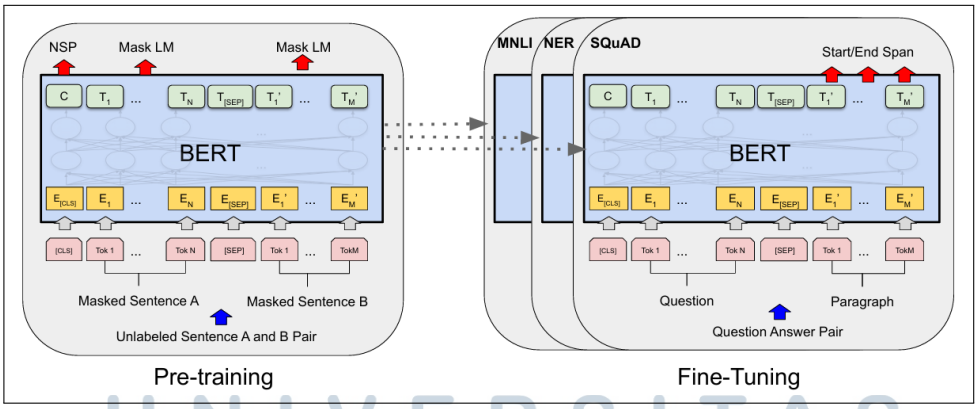
Gambar 2.2. Ilustrasi proses *input* dan *output* BERT

Sumber: [35]

2.8.2 Tahapan Pelatihan Model BERT

Model BERT dirancang melalui dua tahap, yaitu *pre-training* dan *fine-tuning*. Tahap *pre-training* merupakan inisiasi bobot model dengan melatih data tanpa label menggunakan dua metode *unsupervised pre-training*, yaitu *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP). Pada MLM, model belajar memprediksi kata-kata yang disembunyikan dalam sebuah kalimat, sementara NSP melatih model untuk memahami hubungan antar kalimat dengan memprediksi apakah dua kalimat berurutan dalam teks. Proses ini mengubah bobot inisial yang acak menjadi representasi bahasa dasar yang kaya dan kontekstual.

Setelah *pre-training*, model yang telah memiliki bobot dasar akan disesuaikan lebih lanjut. Tahap *fine-tuning* melibatkan pelatihan model secara lebih mendalam dengan menambahkan bobot minimum untuk tugas spesifik (*downstream task*), seperti klasifikasi sentimen, ke seluruh bobot yang sudah di *pre-train*. Hal ini memungkinkan model untuk beradaptasi dalam melakukan tugas yang lebih spesifik dengan data yang relevan. Model ini umumnya terdiri dari 12 layer, 768 *hidden states*, dan 12 *self-attention heads*, yang memungkinkannya memproses dan memahami konteks bahasa secara mendalam dari kedua arah (*bidirectional*). Gambar 2.3 dari proses *pre-training* dan *fine-tuning*.



Gambar 2.3. Ilustrasi proses *pre-training* dan *fine-tuning*

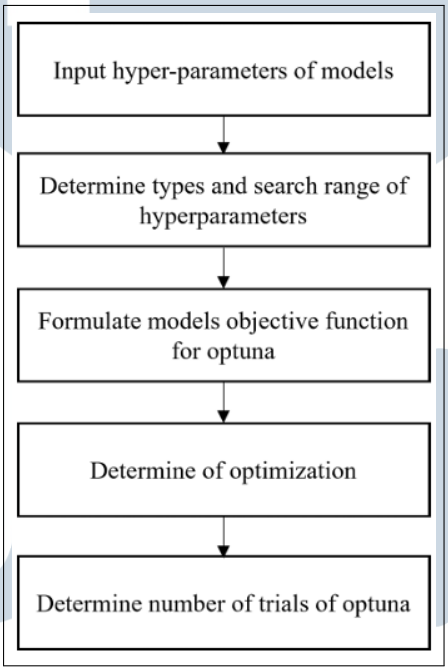
Sumber: [28]

2.8.3 Hyperparameter Tuning

Salah satu tahapan penting dalam mengembangkan dan mengoptimalkan model BERT adalah optimasi *hyperparameter* (*hyperparameter tuning*). Proses ini bertujuan untuk menemukan kombinasi nilai *hyperparameter* terbaik, seperti

learning rate, *batch size*, *weight decay*, dan jumlah *epoch*, untuk meningkatkan performa model. Dalam penelitian ini, metode yang digunakan untuk *hyperparameter tuning* adalah *Bayesian Optimization* dengan bantuan pustaka Optuna [36].

Berbeda dengan metode tradisional seperti *Grid Search* atau *Random Search* yang mengeksplorasi ruang pencarian tanpa mempertimbangkan hasil dari percobaan sebelumnya, *Bayesian Optimization* bekerja dengan membangun model probabilistik dari hasil percobaan sebelumnya. Metode ini, yang sering kali menggunakan pendekatan seperti *Tree-structured Parzen Estimator* (TPE), terbukti lebih efisien dalam menemukan konfigurasi optimal karena mampu mengeksplorasi ruang pencarian yang lebih luas dengan lebih cepat [37]. Prosedur dasar dari proses *hyperparameter tuning* menggunakan pustaka Optuna dapat dilihat pada Gambar 2.4.



Gambar 2.4. Prosedur Optuna untuk *hyperparameter tuning*
Sumber: [38]

Proses ini dimulai dengan Optuna yang secara cerdas menyarankan kombinasi *hyperparameter* baru untuk sebuah percobaan (*trial*). Model kemudian dilatih menggunakan kombinasi parameter tersebut, dan hasilnya dievaluasi berdasarkan sebuah fungsi tujuan (*objective function*), seperti akurasi atau *F1-score*. Optuna akan menyimpan hasil dari setiap *trial* dan menggunakannya untuk menyarankan *hyperparameter* yang lebih menjanjikan untuk *trial* berikutnya. Proses ini terus

berulang hingga mencapai jumlah *trial* yang ditentukan, di mana pada akhirnya akan dipilih kombinasi *hyperparameter* terbaik yang menghasilkan performa paling optimal.

Tujuan utama dari proses *tuning* adalah untuk mencapai keseimbangan antara bias dan variansi, sehingga model dapat berkinerja baik pada data latih dan mampu melakukan generalisasi dengan baik pada data baru.

2.8.4 IndoBERT

IndoBERT merupakan versi turunan dari model BERT yang secara khusus dikembangkan untuk menangani teks berbahasa Indonesia. Berbeda dengan BERT versi multilingual (mBERT), IndoBERT bersifat monolingual karena hanya dilatih menggunakan korpus bahasa Indonesia. Model ini dikembangkan dengan tujuan untuk menghasilkan representasi linguistik yang lebih akurat dan kontekstual dalam bahasa Indonesia, baik formal maupun informal [39]. IndoBERT dilatih menggunakan *dataset* Indo4B, yaitu kumpulan data sebesar ±4 miliar kata dan 250 juta kalimat dari berbagai sumber seperti Wikipedia (74 juta kata), berita dari Kompas, Tempo, Liputan6 (total 55 juta kata), serta Indonesia Web Corpus (90 juta kata) [12][40]. Setelahnya, model dapat disesuaikan lebih lanjut (*fine-tuning*) menggunakan data berlabel untuk berbagai tugas NLP, termasuk klasifikasi sentimen.

Model yang digunakan dalam penelitian ini adalah IndoBERT-base, yang terdiri dari 12 lapisan *Transformer encoder*, 768 dimensi *hidden*, dan 12 *attention heads*, sesuai dengan arsitektur standar BERT-base [11]. Struktur ini mengintegrasikan *multi-head self-attention*, *feedforward neural networks*, serta *layer normalization*, sehingga mampu menangkap relasi antar kata secara *bidirectional* dan memahami konteks kalimat dengan baik.

Pada penelitian ini, IndoBERT digunakan sebagai model utama untuk klasifikasi sentimen komentar Instagram. Proses tokenisasi dilakukan menggunakan *tokenizer* bawaan IndoBERT, yang secara otomatis menambahkan token khusus [CLS] di awal dan [SEP] di akhir teks, serta melakukan *padding* dan *truncation* untuk menyamakan panjang *input*. Hasil tokenisasi kemudian diubah menjadi *tensor* numerik yang diproses oleh model untuk menghasilkan prediksi sentimen berupa label positif, negatif, atau netral.

2.9 Pembagian Dataset (Data Splitting)

Pembagian dataset bertujuan untuk memastikan bahwa model dapat belajar secara efektif, mengoptimalkan kinerjanya, dan dievaluasi secara objektif menggunakan data yang belum pernah dilihat sebelumnya. Hal ini dilakukan untuk mencegah terjadinya *overfitting* maupun *underfitting* [41]. Umumnya, *dataset* dibagi menjadi tiga bagian utama, yaitu data pelatihan (*training set*), data validasi (*validation set*), dan data pengujian (*test set*).

1. Data Pelatihan
   Data pelatihan digunakan untuk melatih model, sehingga model akan belajar pola dan fitur dari data untuk menyesuaikan parameter internal.

2. Data Validasi
   Data validasi digunakan selama proses pelatihan untuk mengevaluasi kinerja model secara berkala dan melakukan penyetelan *hyperparameter* (misalnya, laju pembelajaran, jumlah *epoch*, arsitektur model). Tujuannya adalah untuk membantu mengoptimalkan model tanpa menyentuh data pengujian, sehingga model tidak mengalami *overfitting* terhadap data pelatihan.

3. Data Pengujian
   Data pengujian adalah data yang benar-benar terpisah dan tidak pernah digunakan selama pelatihan maupun validasi. Data pengujian berfungsi sebagai evaluasi akhir yang tidak *bias* terhadap kinerja generalisasi model pada data baru yang belum pernah dilihat. Hasil evaluasi pada data pengujian memberikan gambaran mengenai seberapa baik model dapat bekerja di dunia nyata.

Meskipun tidak ada rasio pembagian yang sempurna karena sangat bergantung pada ukuran *dataset*, kompleksitas masalah, dan model yang digunakan, beberapa rasio telah menjadi praktik umum dan direkomendasikan dalam literatur. Dua rasio yang sering digunakan adalah 70:15:15 dan 80:10:10.

1. Rasio 70:15:15
   Pembagian ini menetapkan 70% data untuk pelatihan, 15% untuk validasi, dan 15% untuk pengujian. Rasio ini sering dipilih ketika ukuran *dataset* cukup besar, memungkinkan model untuk memiliki data yang cukup untuk belajar, sekaligus menyisakan porsi yang layak untuk validasi dan pengujian

yang independen. Ini memberikan keseimbangan yang baik antara data untuk pembelajaran dan data untuk evaluasi kinerja model secara akurat.

2. Rasio 80:10:10
   Rasio ini memberikan porsi yang lebih besar (80%) untuk data pelatihan, sementara menetapkan masing-masing 10% untuk data validasi dan pengujian. Pembagian 80:10:10 sering dipertimbangkan ketika *dataset* relatif lebih kecil, sehingga memaksimalkan data pelatihan menjadi prioritas untuk memastikan model dapat mempelajari pola yang cukup kompleks. Meskipun demikian, tetap diperlukan data validasi dan pengujian yang cukup untuk memastikan evaluasi yang representatif.

Pemilihan antara rasio 70:15:15 dan 80:10:10 sering kali didasarkan pada eksperimen empiris untuk menemukan konfigurasi yang menghasilkan kinerja model terbaik dan paling stabil untuk masalah spesifik yang sedang ditangani.

2.10 Confusion Matrix

Confusion matrix adalah metode yang digunakan untuk menganalisis performa serta mengukur tingkat akurasi suatu model [42]. *Confusion matrix* berbentuk tabel yang menggambarkan jumlah prediksi benar dan salah pada setiap kelas dalam model klasifikasi. Tabel 2.3 menyajikan tampilan *confusion matrix*.

Tabel 2.3. *Confusion matrix*

Kelas	**Predicted Positive**	**Predicted Negative**
Actual Positif	TP (True Positive)	FN (False Negative)
Actual Negatif	FP (False Positive)	TN (True Negative)

Keterangan:

1. TP (*True Positive*) = Data aktual Positif yang diprediksi sebagai Positif
2. TN (*True Negative*) = Data aktual Negatif yang diprediksi sebagai Negatif
3. FP (*False Positive*) = Data aktual Negatif yang diprediksi sebagai Positif
4. FN (*False Negative*) = Data aktual Positif yang diprediksi sebagai Negatif

Selain empat komponen utama, terdapat empat rumus yang digunakan dalam *confusion matrix* utuk perhitungan nilai akurasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score* dengan uraian sebagai berikut:

1. *Accuracy*

   Accuracy merupakan metrik yang digunakan untuk mengukur sejauh mana suatu model dapat memprediksi label secara keseluruhan dengan tepat [42]. Nilai ini dihitung dengan membandingkan jumlah prediksi yang benar dengan total keseluruhan prediksi yang dilakukan. *Accuracy* dapat dihitung dengan menjumlahkan nilai pada diagonal utama, kemudian membaginya dengan total keseluruhan elemen dalam metrik tersebut.

   Rumus 2.2 menunjukkan cara perhitungan *accuracy*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

2. *Precision*

   Precision merupakan metrik yang digunakan untuk mengukur performa seperti sejauh mana prediksi positif yang dibuat oleh model benar-benar relevan [42]. *Precision* didefinisikan sebagai perbandingan antara jumlah prediksi positif yang sebenarnya dengan total keseluruhan prediksi positif [43].

   Rumus 2.3 menunjukkan cara perhitungan *precision*.

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

3. *Recall*

   Recall merupakan metrik yang digunakan untuk mengukur seberapa baik performa suatu model dalam mendeteksi semua sampel positif yang sebenarnya ada dalam data. *Recall* dihitung dengan membagi jumlah positif yang benar dengan total positif yang benar dan positif yang salah [43].

   Rumus 2.4 menunjukkan cara perhitungan *recall*.

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

4. *F1-score*

   F1-score merupakan metrik yang lebih unggul dibandingkan dengan metrik lainnya dalam situasi distribusi kelas yang tidak seimbang [43]. Metrik ini menggabungkan *precision* dan *recall* untuk memberikan gambaran yang lebih baik mengenai performa model.

   Rumus 2.5 menunjukkan cara perhitungan *F1-score*.

$$F1-score = 2 * \frac{precision * recall}{precision + recall} \tag{2.5}$$