BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian digunakan sebagai pembanding antara penelitian yang sudah pernah dilakukan dengan penelitian yang sedang dilakukan oleh penulis, penelitian tersebut adalah sebagai berikut: Menurut Wang, Lu, Chow, dan Zhu (2020) dalam penelitian berjudul COVID-19 sensing: negative sentiment analysis on social media in China via BERT model, permasalahan utama yang diangkat adalah meningkatnya sentimen negatif masyarakat di media sosial selama pandemi COVID-19 yang sulit dianalisis secara akurat karena volume data yang besar dan konteks emosional yang kuat. Penelitian ini menggunakan model BERT untuk mendeteksi sentimen negatif pada unggahan media sosial di China. Hasil penelitian menunjukkan bahwa BERT mampu menangkap konteks linguistik secara mendalam dan memberikan akurasi yang lebih tinggi dibandingkan metode tradisional, khususnya dalam mendeteksi sentimen negatif yang bersifat implisit. Temuan ini menunjukkan keunggulan BERT dalam analisis sentimen pada isu krisis berskala besar (Wang & Lu, 2020).

Menurut Wahyu Widianti dan Saefudin (2025) dalam penelitian berjudul *Analisa sentimen terhadap ulasan pengguna pada aplikasi Polri Presisi menggunakan metode BERT*, permasalahan yang dihadapi adalah beragamnya opini pengguna aplikasi Polri Presisi yang bersifat subjektif, tidak terstruktur, serta banyak menggunakan bahasa informal. Penelitian ini menerapkan metode BERT untuk mengklasifikasikan sentimen

ulasan pengguna menjadi positif, negatif, dan netral. Hasil penelitian menunjukkan bahwa BERT mampu meningkatkan akurasi klasifikasi sentimen secara signifikan dan efektif dalam memahami konteks bahasa Indonesia. Penelitian ini membuktikan bahwa BERT relevan untuk mengevaluasi persepsi publik terhadap layanan digital pemerintah (Widianti & Saefudin, 2025).

Menurut Sholihah, Abdulloh, dan Rahardi (2025) dalam penelitian *Sentiment analysis on KPU performance post-2024 election via YouTube comments using BERT*, permasalahan yang diangkat adalah tingginya dinamika opini publik pasca Pemilu 2024 yang diekspresikan melalui komentar YouTube, dengan banyak komentar bernada emosional, politis, dan kontekstual. Penelitian ini menggunakan model BERT untuk menganalisis sentimen masyarakat terhadap kinerja KPU. Hasil penelitian menunjukkan bahwa BERT mampu mengklasifikasikan sentimen dengan lebih stabil dan akurat meskipun data bersifat polar dan sensitif secara politik. Hal ini menegaskan kemampuan BERT dalam menangani analisis sentimen pada isu demokrasi dan pemilu (Bert et al., 2024).

Selanjutnya, menurut Pratama, Sanjaya, Lubis, Aditya, dan Yennimar (2025) dalam penelitian berjudul *Analisis sentimen publik terkait Danantara menggunakan algoritma IndoBERT pada platform media sosial*, permasalahan yang dihadapi adalah munculnya beragam persepsi publik terhadap isu Danantara yang tersebar di berbagai platform media sosial dengan gaya bahasa yang bervariasi. Penelitian ini menggunakan IndoBERT untuk menyesuaikan karakteristik bahasa Indonesia dalam analisis sentimen. Hasil penelitian menunjukkan bahwa IndoBERT memberikan performa yang lebih baik dibandingkan model non-spesifik bahasa Indonesia, terutama dalam memahami konteks

lokal dan istilah khas. Penelitian ini memperkuat pentingnya penggunaan model BERT yang teradaptasi secara linguistic (Pratama et al., 2025).

Penelitian oleh Liu dan Zhao (2022) dalam studi berjudul *A BERT-based aspect-level sentiment analysis algorithm for cross-domain text* mengangkat permasalahan kesulitan dalam menganalisis sentimen pada level aspek ketika data berasal dari domain yang berbeda. Tantangan utama adalah perbedaan konteks dan distribusi data antar domain yang memengaruhi akurasi model. Penelitian ini mengusulkan pendekatan berbasis BERT untuk analisis sentimen tingkat aspek lintas domain. Hasilnya menunjukkan bahwa BERT mampu mempertahankan performa yang konsisten dan adaptif meskipun diterapkan pada domain yang berbeda. Temuan ini menunjukkan fleksibilitas BERT dalam berbagai konteks analisis sentiment (Liu, 2022).

Dari beberapa penelitian di atas, terlihat bahwa analisis sentimen dengan metode transformer terutama BERT telah terbukti efektif digunakan pada isu-isu publik dan politik pada tahun 2022–2025. Oleh karena itu, penelitian ini akan menggunakan algoritma BERT untuk menganalisis sentimen masyarakat terhadap isu “keaslian ijazah Jokowi”, dengan tujuan menilai kemampuan BERT dalam menangani sarkasme, ironi, ambiguitas, serta bahasa informal, sekaligus mengetahui peningkatannya terhadap akurasi, presisi, recall, dan konsistensi hasil klasifikasi (Pratama et al., 2025).

Tabel 2.1 Tabel Literatur Review (Erna, 20225:9)

No	Judul	Masalah	Tujuan	Metode	Kebaruan	Hasil

1	COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model	Opini publik COVID-19 di media sosial sulit dianalisis manual	Mengidentifikasi sentimen publik	Fine-tuned BERT	Penerapan BERT pada isu krisis kesehatan	BERT mencapai akurasi 75,65% dan mampu mengidentifikasi pola sentimen publik secara kontekstual dibandingkan metode konvensional.
2	Analisis Sentimen Terhadap Ulasan Pengguna Pada Aplikasi Polri Presisi Menggunakan Metode BERT	Ulasan pengguna layanan publik tidak terstruktur	Menganalisis sentimen pengguna	BERT	BERT pada ulasan layanan publik Indonesia	Model mencapai akurasi 86,6% dengan performa baik pada sentimen positif dan negatif, namun kelas netral masih

						kurang optimal.
3	Sentiment Analysis on KPU Performance Post-2024 Election via YouTube Comments Using BERT	Sulit memahami opini publik pasca pemilu	Mengklasifikasikan sentimen komentar	BERT + K-Fold	Analisis sentimen pasca pemilu	BERT menunjukkan akurasi hingga 96% dan efektif memahami konteks opini publik meskipun dipengaruhi distribusi data.
4	Analisis Sentimen Publik Terkait Danantara Menggunakan Algoritma IndoBERT Pada	Opini publik terhadap kebijakan belum terpetakan	Menganalisis sentimen publik	IndoBERT + Augmentasi	IndoBERT khusus Bahasa Indonesia	IndoBERT mencapai akurasi 97,71% dan menunjukkan stabilitas performa pada klasifikasi multi-kelas.

	Platform Media Sosial					
5	A BERT-Based Aspect-Level Sentiment Analysis Algoritma for Cross-Domain Text	Kesulitan analisis sentimen lintas domain	Mengembangkan model lintas domain	BERT + Domain Adversarial	Analisis sentimen level aspek	Model meningkatkan akurasi dan F1-score serta konsisten pada berbagai domain teks.

2.2 Data Mining

Data mining merupakan salah satu bidang penting dalam ilmu komputer dan analisis data yang berfokus pada proses menggali pola, tren, atau informasi yang bermakna dari kumpulan data dalam jumlah besar (Setiawan et al., 2021). Secara etimologis, istilah *data mining* berasal dari dua kata yaitu "data" dan "mining" yang berarti penggalian data. Namun, makna sebenarnya lebih dalam dari sekadar menggali atau mengambil data, karena proses ini melibatkan penerapan algoritma statistik, matematika, dan kecerdasan buatan untuk mengekstraksi pengetahuan tersembunyi (*hidden knowledge*) dari data yang sebelumnya tidak terstruktur atau belum dimanfaatkan secara optimal. Data mining menjadi komponen utama dari proses yang lebih luas yang disebut *Knowledge Discovery in Databases* (KDD), yaitu serangkaian tahapan yang

mengubah data mentah menjadi informasi yang dapat digunakan untuk mendukung pengambilan Keputusan (Marisa et al., 2021).

Proses data mining tidak sekadar mengambil data mentah lalu mengolahnya secara acak, melainkan melalui serangkaian tahapan sistematis. Secara umum, proses KDD terdiri dari lima langkah utama, yaitu: (1) *selection* atau pemilihan data relevan, (2) *preprocessing* atau pembersihan data, (3) *transformation* atau pengubahan data menjadi format yang dapat diproses oleh algoritma, (4) *data mining* atau penerapan metode analisis untuk menemukan pola, dan (5) *interpretation/evaluation* atau tahap evaluasi hasil agar dapat dimaknai secara kontekstual.

2.3 Pengelompokan Data

Pengelompokan data atau *data grouping* merupakan salah satu konsep penting dalam proses analisis data yang bertujuan untuk mengelompokkan sekumpulan data ke dalam kategori atau kelompok tertentu berdasarkan kesamaan karakteristik di antara data tersebut (Hendrastuty., 2024). Dalam konteks ilmu data, pengelompokan dapat dilakukan dengan dua pendekatan utama, yaitu *clustering* (tanpa label atau *unsupervised learning*) dan *classification* (dengan label atau *supervised learning*). *Clustering* berfokus pada menemukan struktur alami dalam data tanpa mengetahui label sebelumnya, sedangkan *classification* digunakan untuk memprediksi label dari data baru berdasarkan model yang telah dilatih. Pada penelitian ini, proses pengelompokan data dilakukan dalam konteks klasifikasi sentimen, di mana data teks dikategorikan ke dalam tiga kelompok utama, yaitu sentimen positif, negatif, dan netral.

Proses pengelompokan data dalam analisis teks dimulai dari tahap representasi data. Teks yang bersifat kualitatif perlu diubah menjadi bentuk numerik agar dapat diproses oleh algoritma komputer. Representasi ini biasanya dilakukan menggunakan metode *feature extraction* seperti *Bag of Words* (BoW), *TF-IDF* (Term Frequency–Inverse Document Frequency), atau representasi vektor kontekstual seperti *word embedding* (Putri & Putra., 2024). Dalam penelitian ini, pendekatan *embedding* yang digunakan berasal dari model BERT, yang mampu menangkap konteks kata dalam kalimat secara dua arah. Hal ini menjadikan pengelompokan data berdasarkan sentimen menjadi lebih akurat, karena model tidak hanya melihat frekuensi kata tetapi juga maknanya dalam konteks tertentu.

Selain itu, pengelompokan data juga berfungsi untuk mengidentifikasi ketidakseimbangan kelas (*class imbalance*). Dalam penelitian berbasis media sosial, fenomena ini umum terjadi karena tidak semua pengguna mengekspresikan opini dengan intensitas yang sama. Biasanya, data berlabel negatif lebih dominan karena topik kontroversial cenderung memicu ekspresi emosional lebih kuat dibandingkan opini positif atau netral. Ketidakseimbangan ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas tertentu, sehingga hasil analisis menjadi kurang akurat. Oleh karena itu, tahap evaluasi distribusi kelas menjadi penting untuk memastikan bahwa proses pengelompokan berjalan secara proporsional dan representatif.

2.4 Proses Data Mining

Proses data mining merupakan tahapan sistematis yang dilakukan untuk mengekstraksi informasi atau pengetahuan berharga dari sekumpulan data yang besar dan

kompleks (Man & Lin, 2021). Data mining tidak berdiri sendiri, melainkan merupakan bagian inti dari rangkaian proses yang lebih luas yang dikenal sebagai *Knowledge Discovery in Database* (KDD). Proses ini bertujuan untuk mengubah data mentah yang semula tidak terstruktur menjadi informasi yang bermakna dan dapat digunakan dalam pengambilan keputusan.

Secara umum, proses data mining terdiri dari beberapa tahapan utama yang saling berhubungan, yaitu:

1. pengumpulan data
2. pemilihan data (*data selection*)
3. pembersihan data (*data cleaning*)
4. transformasi data (*data transformation*)
5. penerapan algoritma penambangan data (*data mining process*)
6. evaluasi serta interpretasi hasil (*pattern evaluation*)

Tahapan-tahapan ini bersifat iteratif, artinya hasil dari satu tahap dapat memengaruhi atau memerlukan perbaikan dari tahap sebelumnya. Dalam penelitian analisis sentimen, setiap tahap memiliki peran yang sangat penting untuk memastikan bahwa hasil akhir yang diperoleh memiliki validitas tinggi dan mencerminkan kondisi sebenarnya di lapangan.

2.5 Machine Learning

Machine Learning atau pembelajaran mesin merupakan salah satu cabang dari kecerdasan buatan *(Artificial Intelligence/AI)* yang berfokus pada pengembangan algoritma yang memungkinkan sistem komputer untuk belajar dari data dan membuat

keputusan atau prediksi tanpa harus diprogram secara eksplisit. Dalam pendekatan konvensional, sistem komputer beroperasi berdasarkan aturan dan logika yang ditentukan oleh manusia (Permadi, 2020). Namun, dalam machine learning, sistem justru belajar mengenali pola dari data yang diberikan dan membangun model matematis untuk melakukan generalisasi terhadap data baru. Konsep dasar ini menjadi pondasi bagi berbagai aplikasi modern, mulai dari pengenalan wajah, analisis sentimen, rekomendasi produk, hingga sistem prediksi cuaca dan deteksi penipuan.

Secara umum, proses machine learning dimulai dengan pengumpulan data, kemudian dilakukan tahap pembersihan dan pra-pemrosesan untuk memastikan kualitas data yang baik (Pradana et al., 2025). Setelah itu, data dibagi menjadi dua bagian utama, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Model machine learning akan dilatih menggunakan data pelatihan untuk mengenali pola yang ada, kemudian diuji menggunakan data pengujian guna mengukur tingkat akurasi dan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Tujuan utamanya adalah agar model mampu memberikan hasil prediksi atau klasifikasi yang mendekati kenyataan di dunia nyata.

Dalam implementasinya, machine learning dibagi menjadi tiga kategori utama berdasarkan cara sistem memperoleh pengetahuan dari data, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Pada *supervised learning*, model dilatih menggunakan dataset yang memiliki label atau target output yang sudah diketahui sebelumnya. Contoh penerapan metode ini adalah klasifikasi sentimen, di mana data teks diberi label positif, negatif, atau netral. Algoritma akan mempelajari hubungan antara teks dengan label tersebut sehingga dapat mengklasifikasikan data baru. Beberapa algoritma

populer dalam kategori ini antara lain *Support Vector Machine (SVM)*, *Decision Tree*, *Naïve Bayes*, dan *Neural Network*.

Sebaliknya, pada *unsupervised learning*, data yang digunakan tidak memiliki label. Tujuannya adalah untuk menemukan struktur atau pola tersembunyi di dalam data tanpa bantuan informasi eksternal. Teknik ini sering digunakan dalam proses pengelompokan (*clustering*) dan reduksi dimensi. Contoh algoritma yang sering digunakan antara lain *K-Means Clustering* dan *Hierarchical Clustering*. Sementara itu, *reinforcement learning* merupakan pendekatan di mana sistem belajar melalui interaksi dengan lingkungan. Model akan mengambil keputusan berdasarkan umpan balik berupa *reward* (hadiah) atau *punishment* (hukuman) dari tindakan yang dilakukan, dan berusaha memaksimalkan nilai kumulatif dari reward yang diperoleh. Pendekatan ini banyak diterapkan pada sistem robotika, game, dan kendaraan otonom.

Seiring dengan perkembangan teknologi dan ketersediaan data yang sangat besar (*big data*), machine learning berkembang menuju tingkat kompleksitas yang lebih tinggi yang dikenal dengan istilah *deep learning* (Rahmawati, 2023). Deep learning merupakan bagian dari machine learning yang menggunakan arsitektur jaringan saraf tiruan (*artificial neural network*) berlapis-lapis untuk mempelajari representasi data secara hierarkis. Model ini mampu mengekstraksi fitur-fitur kompleks dari data mentah, seperti gambar, suara, atau teks, tanpa perlu melakukan ekstraksi fitur manual. Inilah yang kemudian melahirkan algoritma seperti *Convolutional Neural Network (CNN)* untuk data gambar dan *Recurrent Neural Network (RNN)* untuk data berurutan seperti teks atau suara.

2.6 *Deep Learning*

Deep Learning atau pembelajaran mendalam merupakan salah satu cabang lanjutan dari *machine learning* yang meniru cara kerja otak manusia dalam memproses informasi. Konsep dasarnya adalah penggunaan jaringan saraf tiruan (*artificial neural networks*) dengan banyak lapisan (*multiple layers*) yang saling terhubung dan bekerja untuk mengekstraksi pola kompleks dari data (Sabilillah et al., 2024). Jika *machine learning* tradisional masih bergantung pada proses *feature engineering* manual, *deep learning* mampu secara otomatis mengekstraksi dan mempelajari fitur-fitur penting dari data mentah seperti teks, suara, maupun gambar. Pendekatan ini menjadi sangat populer karena kemampuannya menghasilkan performa tinggi dalam berbagai bidang seperti pengenalan citra (*image recognition*), pemrosesan bahasa alami (*natural language processing*), dan analisis sentimen.

Konsep dasar deep learning berakar dari *artificial neural network* (ANN), yang terinspirasi dari struktur neuron biologis pada otak manusia. ANN terdiri dari lapisan neuron buatan yang saling terhubung, yakni *input layer*, *hidden layer*, dan *output layer*. Pada lapisan input, data mentah diterima untuk diproses; lapisan tersembunyi bertugas mengekstraksi pola dan fitur; sedangkan lapisan output menghasilkan prediksi akhir, seperti klasifikasi sentimen positif, negatif, atau netral. Dalam deep learning, jumlah *hidden layer* lebih banyak dibandingkan neural network biasa, sehingga model memiliki kemampuan lebih besar untuk memahami pola yang sangat kompleks dan non-linear.

Salah satu keunggulan utama deep learning adalah kemampuannya dalam *representation learning* — yaitu belajar secara otomatis bagaimana cara terbaik merepresentasikan data agar dapat digunakan untuk tugas tertentu. Misalnya, pada

analisis teks, deep learning dapat mempelajari representasi kata (word embeddings) yang menggambarkan makna kata dalam bentuk vektor numerik. Hal ini membuat model tidak hanya memahami kata secara leksikal, tetapi juga secara kontekstual. Dengan demikian, kata “ijazah” dan “dokumen” misalnya, dapat dianggap memiliki makna yang mirip karena sering muncul dalam konteks yang serupa. Kemampuan ini menjadi dasar keberhasilan model-model modern seperti *Word2Vec*, *GloVe*, dan yang paling mutakhir, *BERT* (Bidirectional Encoder Representations from Transformers).

2.7 *Text Mining*

Text Mining atau penambangan teks adalah salah satu bidang penting dalam ilmu data (*data science*) yang berfokus pada proses ekstraksi informasi dan pengetahuan yang bernilai dari data berbentuk teks tidak terstruktur (Firdaus et al., 2021). Berbeda dengan data numerik yang tersusun rapi dalam tabel, data teks memiliki sifat tidak teratur, beragam, dan kompleks, karena terdiri dari kata, kalimat, serta konteks semantik yang bervariasi. Oleh sebab itu, text mining hadir untuk mengubah teks yang tidak terstruktur menjadi data yang dapat diproses secara komputasional. Proses ini melibatkan serangkaian tahapan seperti pembersihan data (*text preprocessing*), transformasi teks ke dalam bentuk representasi numerik (*vectorization*), analisis pola, serta interpretasi hasil untuk mendapatkan wawasan atau kesimpulan yang bermakna.

Proses text mining umumnya terdiri dari beberapa tahapan penting. Pertama adalah *data collection*, yaitu pengumpulan data teks dari berbagai sumber seperti media sosial, situs berita, atau dokumen resmi. Tahap ini penting karena kualitas hasil analisis sangat bergantung pada relevansi dan volume data yang digunakan. Kedua adalah *text*

preprocessing, di mana data dibersihkan dari elemen-elemen yang tidak relevan seperti tanda baca, angka, emoji, atau kata-kata umum (*stopwords*) yang tidak memiliki makna penting. Ketiga adalah *text transformation*, yaitu proses mengubah teks menjadi representasi numerik agar dapat diproses oleh algoritma pembelajaran mesin. Teknik umum yang digunakan dalam tahap ini antara lain *Bag of Words (BoW)*, *Term Frequency-Inverse Document Frequency (TF-IDF)*, dan *word embeddings* seperti *Word2Vec* atau *BERT embeddings*.

Tahapan berikutnya dalam text mining adalah *pattern discovery* atau pencarian pola. Pada tahap ini, algoritma digunakan untuk menemukan hubungan atau pola tersembunyi dalam data teks. Misalnya, algoritma *clustering* dapat mengelompokkan teks dengan tema yang serupa, sedangkan algoritma *classification* dapat digunakan untuk menentukan kategori sentimen dari suatu teks. Setelah itu, dilakukan *evaluation* dan *interpretation*, yaitu menilai sejauh mana hasil analisis akurat dan bermakna.

2.8 *Text Preprocessing*

Text Preprocessing atau prapemrosesan teks adalah tahap awal yang sangat penting dalam proses *text mining* dan *natural language processing (NLP)*. Tahapan ini bertujuan untuk menyiapkan data teks mentah agar dapat diproses dan dianalisis secara efektif oleh model pembelajaran mesin atau *deep learning* (Tribuana et al., 2025). Data teks yang diperoleh dari media sosial, berita daring, forum, atau sumber digital lainnya sering kali tidak rapi dan mengandung berbagai elemen yang tidak relevan, seperti tanda baca, emotikon, huruf besar-kecil yang tidak konsisten, singkatan, atau kata-kata yang tidak bermakna. Oleh karena itu, proses *text preprocessing* berperan sebagai langkah

pembersihan (*data cleaning*) untuk memastikan bahwa data yang akan digunakan berada dalam format standar, bersih, dan mudah diolah secara komputasional.

Tahapan dalam *text preprocessing* umumnya terdiri dari beberapa langkah berurutan, meskipun urutannya dapat disesuaikan tergantung pada kebutuhan penelitian dan karakteristik data. Secara umum, langkah-langkah tersebut meliputi:

1. Case Folding, Langkah pertama adalah *case folding*, yaitu proses mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase). Hal ini dilakukan untuk menghindari perbedaan perlakuan terhadap kata yang sama namun berbeda penulisan huruf besar-kecilnya. Misalnya, kata "Jokowi" dan "jokowi" dianggap sama setelah proses *case folding*. Langkah ini sederhana tetapi sangat penting agar sistem tidak memperlakukan kata yang sama sebagai entitas berbeda.
2. Tokenization, Setelah semua huruf diubah menjadi huruf kecil, tahap berikutnya adalah *tokenization*, yaitu proses memecah teks menjadi unit-unit yang lebih kecil yang disebut *token*. Token dapat berupa kata, frasa, atau bahkan karakter tergantung pada kebutuhan analisis. Sebagai contoh, kalimat "Isu ijazah palsu Jokowi ramai dibicarakan" akan dipecah menjadi token: ["isu", "ijazah", "palsu", "jokowi", "ramai", "dibicarakan"]. Tahap ini penting agar model dapat memproses setiap kata sebagai satuan informasi yang terpisah.
3. Stopword Removal**,** Stopword adalah kata-kata umum yang sering muncul tetapi tidak memberikan makna signifikan dalam analisis, seperti "yang", "dan", "di", "ke", atau "dengan". Menghapus *stopword* membantu mengurangi kebisingan (*noise*) pada data, sehingga model dapat lebih fokus pada kata-kata yang relevan dengan konteks sentimen. Misalnya, dalam kalimat "Jokowi dituduh memiliki ijazah palsu", kata

"dituduh", "ijazah", dan "palsu" lebih penting daripada kata "memiliki". Penghapusan *stopword* biasanya dilakukan dengan menggunakan daftar kata umum yang sudah ditentukan dalam Bahasa Indonesia.

4. *Stemming dan Lemmatization*, Stemming adalah proses mengubah kata ke bentuk dasarnya dengan menghapus imbuhan seperti awalan dan akhiran. Contohnya, kata "menuduh", "dituduh", dan "menuduhkan" semuanya akan dikembalikan menjadi bentuk dasar "tuduh". Sedangkan *lemmatization* bekerja lebih kompleks karena mempertimbangkan konteks dan struktur gramatikal kata tersebut. Dalam bahasa Indonesia, stemming dapat dilakukan menggunakan algoritma seperti *Nazief-Adriani stemmer* atau pustaka seperti *Sastrawi* yang populer digunakan untuk teks berbahasa Indonesia. Tahap ini membantu model mengenali bahwa kata-kata turunan yang berbeda memiliki akar makna yang sama.
5. *Normalization dan Cleansing*, Normalisasi dilakukan untuk menstandarkan bentuk teks agar konsisten. Misalnya, kata "gk", "ga", dan "nggak" semuanya dinormalisasi menjadi "tidak". Demikian pula, kata "bgt" diubah menjadi "banget". Proses ini sangat penting terutama pada data media sosial yang banyak menggunakan bahasa tidak baku. Selain itu, tahap *cleansing* meliputi penghapusan tanda baca, angka, URL, emoji, mention (@username), dan tagar (#hashtag) yang tidak relevan dengan analisis. Tujuannya agar model tidak terganggu oleh simbol atau elemen non-linguistik.
6. *Handling Negation* dan *Emoji Interpretation*, Dalam analisis sentimen, kata negasi seperti "tidak", "bukan", atau "jangan" memiliki pengaruh besar terhadap makna kalimat. Contohnya, "Jokowi tidak bersalah" memiliki arti berlawanan dengan

"Jokowi bersalah." Oleh karena itu, beberapa penelitian menambahkan tahapan *negation handling* untuk memastikan bahwa kata setelah negasi diberi penanda khusus agar model memahami konteks pembalikan makna. Selain itu, emoji juga dapat memiliki makna emosional, seperti 😡 (marah) atau 😊 (senang), sehingga dalam beberapa kasus, emoji dapat diterjemahkan ke dalam bentuk kata agar tetap memberi kontribusi terhadap analisis sentimen.

Setelah seluruh tahapan *text preprocessing* selesai, teks yang sudah bersih dan terstruktur kemudian diubah menjadi representasi numerik agar dapat diproses oleh model pembelajaran mesin. Teknik representasi ini dikenal sebagai *vectorization*, yang mengubah setiap kata menjadi angka atau vektor. Dalam metode konvensional, digunakan pendekatan seperti *Bag of Words (BoW)* atau *TF-IDF (Term Frequency-Inverse Document Frequency)*. Namun, dalam model modern seperti *BERT*, representasi kata menggunakan konsep *word embeddings*, di mana setiap kata diwakili oleh vektor berdimensi tinggi yang mencerminkan makna semantik dan konteks penggunaannya dalam kalimat.

2.9 Algoritma BERT

Algoritma BERT *(Bidirectional Encoder Representations from Transformers)* merupakan salah satu model *deep learning* yang revolusioner dalam bidang pemrosesan bahasa alami (*Natural Language Processing* atau NLP). Model ini dikembangkan oleh Google AI Research pada tahun 2018 dan telah membawa perubahan besar dalam cara komputer memahami konteks bahasa manusia. Sebelum munculnya BERT, model NLP

umumnya hanya membaca teks dalam satu arah, yaitu dari kiri ke kanan seperti pada model *Recurrent Neural Network (RNN)* dan *Long Short-Term Memory (LSTM)*, atau dari kanan ke kiri seperti pada beberapa model lain. Pendekatan satu arah ini memiliki keterbatasan karena konteks makna suatu kata sering kali dipengaruhi oleh kata-kata di sekitarnya, baik sebelum maupun sesudahnya. Oleh sebab itu, BERT memperkenalkan pendekatan dua arah (*bidirectional*) untuk memahami konteks secara lebih menyeluruh (Septian et al., 2024).

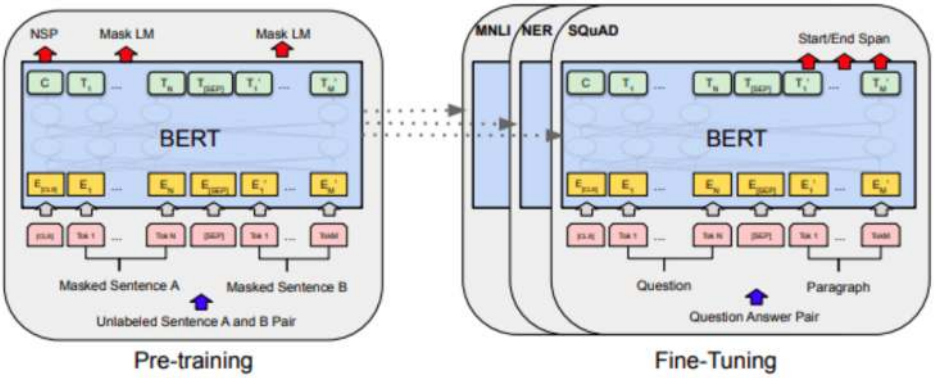
BERT bekerja berdasarkan arsitektur *Transformer*, yaitu model yang menggunakan mekanisme *self-attention* untuk menangkap hubungan antar kata dalam suatu kalimat tanpa bergantung pada urutan kata seperti pada model berbasis urutan (*sequence-based models*). Mekanisme *self-attention* ini memungkinkan BERT untuk menimbang seberapa besar relevansi satu kata terhadap kata lainnya dalam konteks kalimat yang sama. Sebagai contoh, dalam kalimat "Presiden menandatangani ijazah palsu itu bukanlah berita benar," kata "palsu" memiliki arti yang hanya bisa dipahami dengan mempertimbangkan kata-kata "ijazah" dan "berita benar." Model BERT akan memperhatikan hubungan tersebut dari dua arah sekaligus, sehingga mampu menangkap makna yang lebih mendalam dan kontekstual (Srebrovic & Yonamine, 2020).

Arsitektur BERT terdiri dari dua komponen utama, yaitu Encoder dan Decoder, namun dalam implementasinya BERT hanya menggunakan bagian *encoder* dari arsitektur *Transformer*. Encoder berfungsi untuk mengubah teks input menjadi representasi vektor berdimensi tinggi yang mencerminkan makna semantik dari kata-kata tersebut. Dalam konteks ini, setiap kata atau token direpresentasikan sebagai vektor yang

mengandung informasi konteksnya. Semakin tinggi dimensi representasi ini, semakin kaya pula informasi yang dapat ditangkap oleh model.

Proses pelatihan BERT dilakukan melalui dua tahap utama, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP)**.** Pada tahap MLM, sebagian token dalam kalimat akan disembunyikan (*masked*) secara acak, dan model diminta untuk menebak kata yang hilang tersebut berdasarkan konteks sekitarnya. Contohnya, jika kalimat "Presiden Jokowi menandatangani [MASK] palsu" diberikan, model akan berusaha memprediksi kata yang hilang, misalnya "ijazah," dengan memperhatikan hubungan antar kata yang tersisa. Tahap ini membantu model memahami hubungan semantik antar kata. Sementara itu, pada tahap NSP, model dilatih untuk memprediksi apakah dua kalimat saling berhubungan atau tidak. Tahapan ini penting agar BERT dapat memahami hubungan antarkalimat, seperti dalam analisis teks panjang atau percakapan (Sriyanti et al., 2024).

Salah satu keunggulan utama BERT adalah kemampuannya untuk transfer learning**,** yaitu model yang sudah dilatih sebelumnya (*pre-trained*) pada dataset besar seperti Wikipedia dapat disesuaikan (*fine-tuned*) untuk berbagai tugas NLP tertentu seperti klasifikasi sentimen, analisis topik, *named entity recognition*, dan *question answering*. Dalam konteks penelitian "Analisis Sentimen Masyarakat terhadap Isu Ijazah Palsu Jokowi menggunakan Algoritma BERT", model BERT yang telah dilatih pada bahasa Indonesia, seperti *IndoBERT*, digunakan untuk memahami konteks teks dalam Bahasa Indonesia yang sering kali mengandung variasi bahasa, gaya informal, serta campuran kata baku dan tidak baku sebagaimana yang ditemukan pada media sosial.



Gambar 2.1 Pre-training dan Fine-tuning pada BERT (Ashour Ali, 2023:26)

BERT memiliki dua kerangka kerja utama dengan fungsi yang berbeda, yaitu *framework pre-training* dan *framework fine-tuning*. Pada framework pre-training (ditunjukkan pada Gambar 2.1), BERT dilatih menggunakan data besar tanpa label melalui berbagai tugas pre-training. Sementara itu, pada framework fine-tuning, model BERT diinisialisasi menggunakan parameter hasil pre-training dan kemudian dilatih ulang menggunakan data berlabel sesuai dengan kebutuhan *downstream task*. Setiap *downstream task* memiliki model fine-tuning yang berbeda, meskipun semuanya berasal dari model pre-trained yang sama. Untuk mempermudah pemahaman, dapat dicontohkan bagaimana BERT digunakan dalam menjawab pertanyaan. Beragam fitur BERT dapat dikombinasikan untuk menangani berbagai jenis tugas, dengan hanya terdapat sedikit perbedaan antara arsitektur pre-trained dan arsitektur downstream.

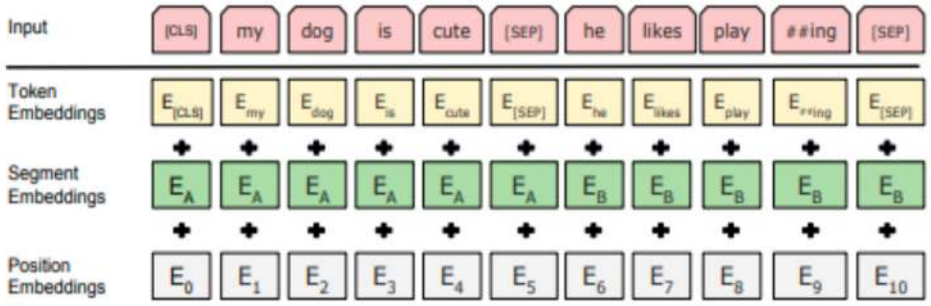
a. Arsitektur Model, Arsitektur BERT menggunakan encoder transformer dua arah, yang mengacu pada implementasi yang dijelaskan pada [18] dan dipublikasikan

melalui library Tensor2Tensor. Saat ini, transformer telah menjadi arsitektur yang sangat umum digunakan, dan penerapannya pada BERT mengikuti prinsip dasar transformer pada umumnya.

b. Representasi Input/Output, Agar BERT dapat menangani berbagai downstream task, representasi input dirancang untuk dapat merepresentasikan satu kalimat maupun pasangan kalimat (misalnya pasangan pertanyaan–jawaban) dalam satu urutan token. Dalam konteks ini, istilah "kalimat" tidak selalu merujuk pada kalimat linguistik yang utuh, melainkan dapat berupa potongan teks berurutan. Sementara itu, "urutan" mengacu pada rangkaian token yang menjadi input BERT, baik berupa satu kalimat tunggal maupun dua kalimat yang digabungkan. BERT menggunakan WordPiece embeddings dengan kosakata sebanyak 30.000 token. Token pertama pada setiap urutan selalu berupa token khusus [CLS], yang representasi tersembunyinya pada layer terakhir digunakan sebagai representasi agregat untuk tugas klasifikasi. Untuk pasangan kalimat, keduanya digabungkan dalam satu urutan dan dibedakan dengan dua mekanisme, yaitu penggunaan token khusus [SEP] sebagai pemisah, serta penambahan segment embedding yang menunjukkan apakah suatu token berasal dari kalimat A atau kalimat B. Seperti yang ditunjukkan pada Gambar 1, embedding input dilambangkan dengan E, vektor tersembunyi terakhir dari token [CLS] dilambangkan sebagai $C \in \mathbb{R}^H$, dan vektor tersembunyi terakhir untuk token input ke-i dilambangkan sebagai $T_i \in \mathbb{R}^H$. Representasi input untuk setiap token dibentuk dengan menjumlahkan token embedding, segment embedding, dan positional embedding yang sesuai. Visualisasi proses pembentukan representasi ini ditampilkan pada Gambar 2.2.

c. Pre Trained BERT, Pada tahap pre-training, BERT menggunakan dua jenis tugas unsupervised yang dijelaskan sebagai berikut.

1. *Masked Language Model*, secara konsep, model dua arah yang dalam diyakini memiliki kemampuan yang lebih baik dibandingkan model satu arah kiri-ke-kanan, kanan-ke-kiri, maupun kombinasi dangkal dari keduanya.



Gambar 2.2 Konstruksi Input/Output BERT (Moksh Shukla, 2023:28)

Namun, model bahasa bersyarat konvensional hanya dapat dilatih secara satu arah, karena pelatihan dua arah memungkinkan setiap kata memperoleh informasi tentang dirinya sendiri secara tidak langsung, sehingga model dapat dengan mudah menebak kata target. Untuk mengatasi permasalahan tersebut dan memungkinkan pembelajaran representasi dua arah yang dalam, BERT menerapkan pendekatan dengan menutupi sebagian token input secara acak, kemudian memprediksi token yang disamarkan tersebut. Proses ini disebut Masked Language Model (MLM), yang dalam literatur juga dikenal sebagai tugas Cloze. Pada mekanisme ini, vektor tersembunyi terakhir dari token yang dimask digunakan sebagai masukan ke lapisan softmax untuk memprediksi

kata asli berdasarkan kosakata, serupa dengan language model standar. Dalam setiap urutan input, sebanyak 15% token WordPiece dipilih secara acak untuk dilakukan masking. Berbeda dengan pendekatan denoising auto-encoder, BERT hanya memprediksi token yang disamarkan, bukan merekonstruksi seluruh input. Pendekatan ini memang memungkinkan pembelajaran dua arah, tetapi menimbulkan perbedaan antara tahap pre-training dan fine-tuning, karena token [MASK] tidak digunakan pada saat fine-tuning. Untuk meminimalkan perbedaan tersebut, token yang terpilih tidak selalu diganti dengan [MASK]. Dari 15% token yang dipilih, 80% diganti dengan token [MASK], 10% diganti dengan token acak, dan 10% dibiarkan tetap sama. Selanjutnya, representasi tersembunyi Ti digunakan untuk memprediksi token asli dengan fungsi loss berupa cross-entropy.

2. *Next Sentence Prediction*, Banyak downstream task penting, seperti *Question Answering* (QA) dan Natural Language Inference (NLI), bergantung pada kemampuan model dalam memahami hubungan antar dua kalimat. Hubungan ini tidak dapat ditangkap secara langsung hanya melalui pemodelan bahasa. Untuk melatih model agar memahami relasi antar kalimat, BERT menggunakan tugas pre-training berupa prediksi kalimat berikutnya atau *Next Sentence Prediction* (NSP). Dalam setiap contoh pre-training, pasangan kalimat A dan B dipilih dengan dua kemungkinan: pada 50% kasus, B merupakan kalimat yang benar-benar muncul setelah A dalam korpus dan diberi label IsNext, sedangkan pada 50% kasus lainnya, B merupakan kalimat acak dari korpus dan diberi label NotNext. Seperti ditunjukkan pada Gambar

1, representasi dari token [CLS] digunakan untuk melakukan prediksi NSP. Meskipun sederhana, tugas ini terbukti sangat membantu performa BERT pada tugas QA dan NLI. Tugas NSP memiliki keterkaitan dengan tujuan pembelajaran representasi pada penelitian sebelumnya, namun perbedaannya terletak pada pendekatan transfer pembelajaran. Pada BERT, seluruh parameter hasil pre-training ditransfer untuk menginisialisasi model pada end-task, bukan hanya embedding kalimat.

d. Fine Tuning BERT, Proses fine-tuning pada BERT relatif sederhana karena arsitektur transformer dengan mekanisme self-attention memungkinkan penggunaan satu model untuk berbagai downstream task. Pada pendekatan tradisional yang melibatkan pasangan teks, biasanya setiap teks dikodekan secara terpisah sebelum diterapkan mekanisme cross-attention dua arah. BERT menyatukan kedua tahap tersebut dengan mengodekan pasangan teks secara langsung menggunakan self-attention, sehingga memungkinkan terjadinya bidirectional cross-attention antar dua kalimat dalam satu proses. Untuk setiap task, pengguna hanya perlu menambahkan komponen input dan output yang spesifik terhadap task tersebut, kemudian melakukan fine-tuning seluruh parameter model secara end-to-end. Pada sisi input, kalimat A dan B yang digunakan saat pre-training dianalogikan sebagai pasangan kalimat pada berbagai skenario, seperti pasangan kalimat pada tugas parafrase, pasangan hipotesis–premis pada entailment, pasangan konteks–pertanyaan pada question answering, serta pasangan teks pada klasifikasi teks atau penandaan urutan. Pada sisi output, representasi token digunakan sebagai masukan ke lapisan keluaran untuk tugas

tingkat token, seperti sequence labeling atau question answering, sedangkan representasi token [CLS] digunakan untuk tugas klasifikasi, seperti analisis sentimen atau entailment. Dibandingkan tahap pre-training, proses fine-tuning memerlukan biaya komputasi yang relatif lebih rendah.

e. IndonLU, IndoNLU merupakan salah satu inisiatif yang dikembangkan oleh komunitas IndoBenchmark. IndoNLU Benchmark menyediakan kumpulan sumber daya yang digunakan untuk melatih, mengevaluasi, dan menganalisis sistem pemahaman bahasa alami dalam bahasa Indonesia. IndoNLU mencakup dua belas jenis task, mulai dari klasifikasi kalimat tunggal hingga pelabelan urutan pada pasangan kalimat dengan tingkat kompleksitas yang bervariasi. Dataset yang digunakan berasal dari berbagai domain dan gaya bahasa, sehingga menjamin keberagaman task yang dievaluasi.

2.10 Platform X

Platform X, yang sebelumnya dikenal sebagai Twitter, merupakan salah satu media sosial paling berpengaruh di dunia dalam hal penyebaran informasi dan pembentukan opini publik. Sejak diakuisisi dan direbranding oleh Elon Musk pada tahun 2023, platform ini mengalami sejumlah perubahan signifikan, baik dari sisi fitur, kebijakan, maupun arah strateginya. Meskipun demikian, esensi utama Platform X tetap sama: menjadi ruang digital di mana pengguna dapat mengekspresikan pendapat, berbagi berita, serta berinteraksi secara langsung dengan individu lain, lembaga, dan tokoh publik dalam bentuk pesan singkat yang disebut *tweet*.

Platform X memiliki karakteristik unik yang menjadikannya berbeda dari media sosial lainnya seperti Facebook atau Instagram. Batasan jumlah karakter pada setiap *tweet* (280 karakter) mendorong pengguna untuk mengekspresikan opini mereka secara padat dan langsung ke inti permasalahan. Sifat singkat ini membuat data yang dihasilkan relatif mudah untuk diolah dalam analisis teks, termasuk analisis sentimen. Selain itu, fitur *hashtag* (#) memungkinkan pengelompokan topik tertentu sehingga percakapan publik tentang isu yang sama dapat dengan mudah dilacak. Misalnya, dalam isu "ijazah palsu Jokowi", berbagai *hashtag* seperti **#**ijazahpalsu**,** #Jokowi**,** atau #ijazahgate sering digunakan oleh pengguna untuk menandai opini mereka. Hal ini membuat peneliti dapat mengumpulkan data dengan lebih terarah dan kontekstual.