

BAB 2

LANDASAN TEORI

Bab ini menyajikan dasar-dasar teori yang mendasari penelitian mengenai analisis sentimen politik dengan menggunakan model *IndoBERT* dan integrasi *generative AI* untuk menghasilkan penjelasan naratif. Penjelasan teori dimulai dari konsep dasar analisis sentimen, *platform X* sebagai sumber data, model bahasa berbasis *transformer*, serta peran *generative AI* dalam interpretasi hasil.

2.1 Analisis Sentimen

Analisis sentimen adalah proses komputasional yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini dalam teks untuk menentukan sikap penulis terhadap suatu topik, entitas, atau peristiwa [21]. Dalam konteks politik, analisis sentimen memungkinkan peneliti untuk memahami opini publik terhadap kandidat, kebijakan, atau isu-isu tertentu melalui data yang dikumpulkan dari platform digital [22].

Pendekatan analisis sentimen telah berkembang dari metode berbasis leksikon ke metode pembelajaran mesin yang lebih canggih. Pendekatan berbasis leksikon mengandalkan kamus kata dengan nilai sentimen yang telah ditentukan, yang kemudian digunakan untuk menghitung skor sentimen keseluruhan dari teks [23]. Meskipun relatif sederhana dalam implementasinya, pendekatan ini memiliki keterbatasan dalam menangani konteks, sarkasme, dan nuansa bahasa.

Metode pembelajaran mesin, terutama pendekatan *deep learning*, menawarkan peningkatan signifikan dalam akurasi analisis sentimen dengan kemampuannya untuk mempelajari fitur-fitur kompleks dari data. Model-model ini dapat mengidentifikasi pola-pola yang halus dalam teks yang mungkin tidak tertangkap oleh pendekatan berbasis aturan [24].

Penerapan analisis sentimen dalam konteks pemilu Indonesia telah menunjukkan potensi untuk melengkapi metode survei tradisional. Beberapa penelitian sebelumnya telah menggunakan analisis sentimen untuk memetakan opini publik terhadap kandidat politik pada saat pemilihan kepala daerah, menunjukkan adanya korelasi antara sentimen *online* dan preferensi pemilih [25].

2.2 Platform X (Twitter)

Platform X, sebelumnya dikenal sebagai Twitter, adalah media sosial berbasis *microblogging* yang memungkinkan pengguna untuk membagikan pesan singkat atau *tweet* dengan batasan 280 karakter [1]. Platform ini dilengkapi dengan fitur-fitur seperti *hashtag*, *mention*, dan *retweet* yang memfasilitasi penyebaran informasi dan diskusi publik [2].

Di Indonesia, Platform X menjadi salah satu ruang digital utama untuk diskusi politik, terutama menjelang pemilihan umum. Politisi, aktivis, jurnalis, dan masyarakat umum menggunakan platform ini untuk berbagi pendapat, mendiskusikan isu-isu politik, dan merespon peristiwa terkini [26]. Karakteristik platform yang memungkinkan percakapan lintas kelompok sosial dan demografis menjadikannya sumber data yang berharga untuk analisis sentimen politik.

Meskipun jumlah pengguna Platform X di Indonesia tidak sebanyak platform media sosial lainnya, pengaruhnya dalam diskusi politik cukup signifikan. Hal ini karena sifat platform yang lebih terbuka dan berorientasi pada diskusi publik dibandingkan platform lain yang cenderung lebih personal [27]. Selama periode pemilu, volume percakapan politik di Platform X meningkat tajam, memberikan *dataset* yang kaya untuk analisis sentimen.

2.3 Text Preprocessing

Text preprocessing adalah tahap awal dalam analisis sentimen untuk membersihkan dan mempersiapkan data teks agar sesuai dengan model pembelajaran mesin. Dalam penelitian ini, data teks berasal dari tweet di Platform X yang dikumpulkan pada Oktober 2022 hingga April 2023, dengan ciri khas seperti batasan karakter, *hashtag*, *mention*, dan bahasa *informal* [1].

Proses ini mencakup beberapa langkah sederhana. Pertama, *tokenization* memecah teks menjadi kata-kata atau frasa. Kedua, *normalization* mengubah teks menjadi bentuk standar, seperti huruf kecil dan perbaikan *slang* (misalnya, "bgs" menjadi "bagus"). Ketiga, *stopword removal* menghapus kata umum seperti "dan" atau "yang" menggunakan daftar *stopwords* bahasa Indonesia [28].

Langkah lain meliputi penghapusan *noise* seperti *mention* (@username), *hashtag* (#topik), dan URL, serta penanganan *emojis* dan *slang* yang relevan dengan sentimen politik. Terakhir, *stemming* dengan algoritma seperti Sastrawi digunakan untuk mengubah kata ke bentuk dasar, seperti "berlari" menjadi "lari", agar model

IndoBERT dapat memproses teks secara lebih baik.

2.4 *Transformer*

Transformer adalah arsitektur *deep learning* yang diperkenalkan oleh Vaswani et al. pada tahun 2017 dalam makalah berjudul *Attention is All You Need* [29]. Arsitektur ini menggantikan pendekatan berbasis *Recurrent Neural Networks* (*RNN*) yang sebelumnya dominan dalam pemrosesan bahasa alami dengan mekanisme *self-attention*, yang memungkinkan model untuk mempertimbangkan semua kata dalam kalimat secara bersamaan.

Komponen utama dari *Transformer* adalah mekanisme *attention* yang memungkinkan model untuk memberikan bobot yang berbeda pada bagian-bagian input yang berbeda saat memproses sekuens.

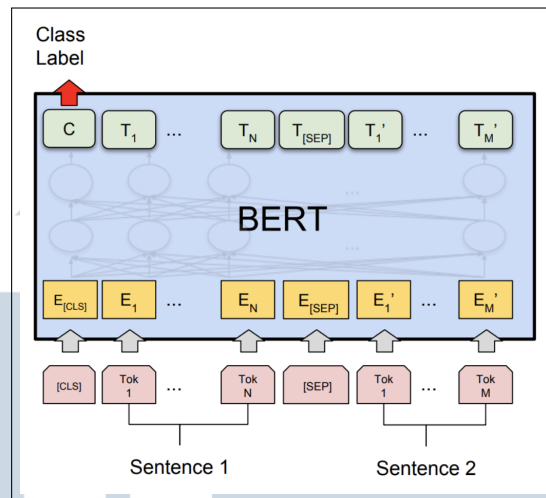
Keunggulan *Transformer* dibandingkan arsitektur sebelumnya terletak pada kemampuannya untuk memproses data secara paralel, yang signifikan mengurangi waktu pelatihan dan memungkinkan model untuk dilatih pada *dataset* yang jauh lebih besar. Selain itu, *Transformer* lebih efektif dalam menangkap dependensi jarak jauh dalam teks, yang penting untuk pemahaman bahasa alami [30].

2.5 *BERT (Bidirectional Encoder Representations from Transformers)*

BERT adalah model bahasa berbasis *Transformer* yang dikembangkan oleh Google pada 2018 [11]. Inovasi utama BERT adalah penggunaan *pre-training* bidirectional yang memungkinkan model untuk mempertimbangkan konteks kata dari kedua arah, tidak seperti model unidirectional sebelumnya yang hanya mempertimbangkan konteks dari satu arah.

Pre-training BERT melibatkan dua tugas utama: *Masked Language Modeling* (*MLM*) dan *Next Sentence Prediction* (*NSP*). Dalam *MLM*, beberapa *token* input secara acak dimasking, dan model dilatih untuk memprediksi *token* yang dimasking berdasarkan konteksnya. Dalam *NSP*, model dilatih untuk memprediksi apakah dua kalimat yang diberikan secara berurutan dalam teks asli atau tidak [11].

Setelah *pre-training*, BERT dapat *fine-tuned* pada berbagai tugas NLP dengan menambahkan lapisan output dan melatih seluruh model pada *dataset* berlabel untuk tugas spesifik. Pendekatan *pre-training* dan *fine-tuning* ini telah terbukti sangat efektif, dengan BERT mencapai hasil *state-of-the-art* pada berbagai *benchmark* NLP [11].



Gambar 2.1. Arsitektur BERT untuk Tugas Klasifikasi Teks
Sumber: [31]

Pada Gambar 2.1, ditunjukkan proses di mana dua kalimat (*Sentence 1* dan *Sentence 2*) dipisahkan oleh token [SEP], sementara token [CLS] di awal digunakan sebagai representasi keseluruhan untuk tujuan klasifikasi. Setiap token melewati tahap embedding (token, segment, dan position), kemudian diproses melalui beberapa lapisan *Transformer encoder*. Keluaran akhir dari token [CLS] menjadi masukan lapisan klasifikasi untuk menghasilkan *class label*.

Melalui mekanisme *self-attention*, BERT dapat memahami konteks kata secara lebih menyeluruh, baik dari kiri maupun kanan token. Pendekatan ini membuat BERT unggul dalam berbagai tugas pemrosesan bahasa alami, termasuk klasifikasi teks, analisis sentimen, *question answering*, dan lain-lain.

2.6 IndoBERT

IndoBERT adalah adaptasi model BERT yang dilatih khusus untuk bahasa Indonesia [32]. Model ini dikembangkan untuk mengatasi keterbatasan model multibahasa seperti mBERT dalam menangani keunikan bahasa Indonesia, termasuk variasi dialek, struktur sintaksis, dan penggunaan kata yang khas [28].

Pre-training IndoBERT dilakukan pada korpus bahasa Indonesia yang mencakup berbagai sumber seperti artikel berita, Wikipedia Indonesia, dan konten media sosial. Keragaman sumber ini memungkinkan model untuk memahami berbagai *register* bahasa Indonesia, dari formal hingga informal [28].

IndoBERT telah menunjukkan peningkatan performa yang signifikan dibandingkan model bahasa umum pada berbagai tugas NLP bahasa Indonesia,

termasuk analisis sentimen. Evaluasi pada *benchmark* IndoNLU menunjukkan bahwa IndoBERT secara konsisten mengungguli model multibahasa pada berbagai tugas pemrosesan bahasa Indonesia [32].

Dalam konteks analisis sentimen politik, IndoBERT menawarkan keunggulan karena kemampuannya untuk menangkap nuansa bahasa Indonesia yang digunakan dalam diskusi politik, termasuk idiom, referensi budaya, dan konteks lokal yang mungkin tidak dipahami dengan baik oleh model yang dilatih pada korpus multibahasa [25].

2.7 *Fine-tuning* Model Bahasa

Fine-tuning adalah proses adaptasi model bahasa *pre-trained* untuk tugas spesifik menggunakan dataset yang lebih kecil [33]. Proses ini memanfaatkan *transfer learning*, di mana pengetahuan yang diperoleh model selama *pre-training* ditransfer dan disesuaikan untuk tugas baru.

Dalam konteks penelitian ini, *fine-tuning* IndoBERT untuk analisis sentimen *tweet* politik melibatkan beberapa tahapan. Pertama, model IndoBERT *pre-trained* dimodifikasi dengan menambahkan lapisan klasifikasi untuk tugas sentimen biner (positif/negatif). Kemudian, model dilatih pada *dataset tweet* politik berlabel dengan menggunakan *learning rate* yang lebih kecil dibandingkan *pre-training* untuk menjaga pengetahuan bahasa yang telah diperoleh sebelumnya [34].

Beberapa pertimbangan penting dalam *fine-tuning* termasuk pemilihan *hyperparameter* seperti *learning rate*, *batch size*, dan jumlah *epoch* pelatihan. Implementasi teknik regularisasi seperti *dropout* dan *weight decay* juga penting untuk mencegah *overfitting*, terutama ketika *dataset fine-tuning* relatif kecil [34].

Penelitian sebelumnya menunjukkan bahwa *fine-tuning* IndoBERT pada *dataset tweet* bahasa Indonesia dapat meningkatkan akurasi analisis sentimen secara signifikan dibandingkan dengan model tradisional seperti SVM atau LSTM [35].

2.8 *Generative AI* untuk Penjelasan Naratif

Generative AI merujuk pada teknologi kecerdasan buatan yang mampu menghasilkan konten baru seperti teks, gambar, atau audio berdasarkan pola yang dipelajari dari data pelatihan [36]. Dalam konteks analisis data, *generative AI* dapat digunakan untuk mengubah hasil kuantitatif menjadi narasi yang bermakna dan mudah dipahami.

Model bahasa generatif seperti GPT (*Generative Pre-trained Transformer*) menggunakan arsitektur *transformer* yang dilatih untuk memprediksi *token* berikutnya dalam urutan teks berdasarkan konteks sebelumnya. Pelatihan pada korpus teks yang sangat besar memungkinkan model ini untuk menghasilkan teks yang koheren dan kontekstual untuk berbagai aplikasi [37].

Penjelasan naratif yang dihasilkan oleh *generative AI* memiliki beberapa keunggulan dibandingkan visualisasi tradisional untuk interpretasi hasil analisis sentimen. Pertama, narasi dapat menyampaikan konteks dan latar belakang yang mungkin tidak terlihat dalam grafik atau tabel. Kedua, narasi dapat menjelaskan hubungan sebab-akibat antara peristiwa politik dan perubahan sentimen. Ketiga, narasi dapat disesuaikan untuk berbagai audiens dengan tingkat pemahaman teknis yang berbeda [38].

Dalam penelitian ini, *generative AI* digunakan untuk menghasilkan penjelasan naratif dari hasil analisis sentimen, memungkinkan interpretasi yang lebih menyeluruh dan kontekstual tentang opini publik terhadap kandidat presiden di Platform X. Pendekatan ini membantu menjembatani kesenjangan antara analisis data kuantitatif dan pemahaman kualitatif yang diperlukan untuk wawasan yang berarti dalam konteks politik.

2.9 Pemilu Presiden Indonesia 2024

Pemilihan Umum (Pemilu) Presiden Indonesia 2024 dijadwalkan berlangsung pada 14 Februari 2024, diselenggarakan oleh Komisi Pemilihan Umum (KPU) [39]. Pemilu ini memiliki signifikansi khusus karena menandai transisi kepemimpinan setelah dua periode pemerintahan Presiden Joko Widodo.

Proses pemilu di Indonesia dilaksanakan melalui serangkaian tahapan yang diatur dalam UU No. 7 Tahun 2017 tentang Pemilihan Umum. Tahapan ini meliputi pendaftaran partai politik, penetapan peserta pemilu, pencalonan presiden dan wakil presiden, masa kampanye, pemungutan suara, dan penetapan hasil [40].

Periode pra-pemilu, yang menjadi fokus penelitian ini, merupakan fase penting di mana kandidat mempresentasikan visi dan program mereka, sementara pemilih mulai mempertimbangkan pilihan politik. Selama periode ini, diskusi politik di media sosial, termasuk Platform X, meningkat secara signifikan, mencerminkan dinamika opini publik dan sentimen terhadap kandidat [41].

Digitalisasi kampanye politik menjadi tren yang semakin menonjol dalam pemilu Indonesia, dengan kandidat dan partai memanfaatkan media sosial untuk

menjangkau pemilih. Studi menunjukkan bahwa sebagian besar pemilih Indonesia, terutama generasi muda, menggunakan media sosial sebagai sumber informasi politik utama [42].

Analisis sentimen media sosial dalam konteks Pemilu 2024 dapat memberikan wawasan berharga tentang persepsi publik terhadap kandidat, efektivitas strategi kampanye, dan isu-isu yang paling menarik perhatian pemilih. Pendekatan ini melengkapi metode penelitian opini publik tradisional dengan menawarkan data sentimen *real-time* yang dapat membantu pemahaman yang lebih komprehensif tentang dinamika politik pra-pemilu.

