BAB II
TINJAUAN PUSTAKA

2.1 Studi Terkait

Penelitian yang dilakukan oleh Raden Mas Rizqi et al.[3] menemukan bahwa model BERT mampu mencapai akurasi hingga 99% dalam analisis sentimen ulasan aplikasi Ruangguru di Google Play Store. Studi tersebut juga menegaskan bahwa BERT memiliki keunggulan dibandingkan model tradisional seperti Support Vector Machine (SVM) dan Naive Bayes dalam menangani teks berbahasa Indonesia. Selanjutnya, kajian yang dilakukan oleh Sumayah S et al.[7] membandingkan performa BERT dan SVM dalam analisis sentimen terhadap topik Metaverse. Hasil penelitian tersebut menunjukkan bahwa BERT mencapai akurasi lebih tinggi (94%) dibandingkan dengan SVM (81%), menegaskan bahwa BERT lebih efektif dalam menangani teks dengan variasi bahasa dan ekspresi emosional yang kompleks. Sementara itu, penelitian lain yang dilakukan oleh Hanvinto Maichael et al.[8] membandingkan performa metode SVM dengan BERT pada analisis sentimen menggunakan dataset dari Twitter yang diperoleh melalui teknik crawling. Hasilnya menunjukkan bahwa algoritma BERT memiliki performa yang lebih unggul dengan akurasi sebesar 94%, sedikit lebih tinggi dibandingkan algoritma SVM yang memperoleh akurasi 93%.

2.2 Analisis Sentimen

Analisis sentimen merupakan proses yang mencakup pengumpulan, ekstraksi, dan pengidentifikasian sentimen atau opini yang terkandung dalam suatu teks. Proses ini bertujuan untuk memahami emosi atau pandangan yang berkaitan dengan entitas tertentu, topik, maupun peristiwa tertentu [9]. Analisis sentimen merupakan metode yang efektif untuk memahami respons publik terhadap acara pendidikan, seperti “Clash of Champions”. Sentimen yang diungkapkan baik positif, negatif, maupun netral ataupun *“others”* melalui komentar masyarakat dapat memberikan wawasan mengenai penerimaan acara tersebut di kalangan masyarakat. Hal ini khususnya terkait dengan potensi peningkatan minat belajar yang menjadi salah satu tujuan utama dari penyelenggaraan acara tersebut. Penelitian yang dilakukan oleh Helmiyah S, dkk[10] menunjukkan jika analisis sentimen yang dilakukan di media sosial dapat memberikan wawasan yang mendalam mengenai presepsi masyarakat terhadap program pendidikan berbasis online.

2.3 *Natural Language Processing* (NLP) dalam Analisis Sentimen

Pemrosesan Bahasa Alami (*Natural Language Processing*/NLP) adalah salah satu cabang kecerdasan buatan yang bertujuan untuk mengolah interaksi antara manusia dan mesin melalui penggunaan bahasa alami. Dalam konteks analisis sentimen, NLP digunakan untuk mengekstraksi, menganalisis, dan mengklasifikasikan informasi dari teks yang bersifat tidak terstruktur. Salah satu model NLP terkini, yaitu BERT (*Bidirectional Encoder Representations from Transformers*), telah menunjukkan keunggulan signifikan dalam berbagai tugas analisis sentimen. Keunggulan tersebut didasarkan pada kemampuannya untuk memahami konteks secara mendalam dan komprehensif [11]. Analisis sentimen sendiri memiliki keterkaitan yang erat dengan *Natural Language Processing* (NLP) karena juga memanfaatkan berbagai teknik pemrosesan NLP yang bertujuan untuk memahami serta mengklasifikasikan teks berdasarkan ekspresi emosional yang terkandung di dalamnya. Selain itu, penerapan model pembelajaran mesin berpotensi meningkatkan kemampuan prediksi sentimen dengan tingkat akurasi yang lebih tinggi.

2.4 Algoritma BERT (*Bidirectional Encoder Representations from Transformers*)

Penggunaan model Bidirectional Encoder Representations from Transformers (BERT) dalam penelitian ini didasarkan pada pertimbangan teoritis dan empiris yang relevan dengan perkembangan terkini di bidang Pemrosesan Bahasa Alami (Natural Language Processing/NLP). Model BERT, yang diperkenalkan pada tahun 2018 oleh Devlin et al.[12], telah menunjukkan keunggulannya dalam berbagai tugas NLP berkat kemampuannya dalam memahami konteks secara dua arah (bidirectional context). Berbeda dengan model terdahulu seperti Recurrent Neural Network (RNN) dan Long Short-Term Memory (LSTM) yang memproses teks secara sekuensial, BERT mengadopsi arsitektur Transformer yang memungkinkan pemrosesan paralel. Keunggulan ini membuat BERT lebih efektif dalam menangkap hubungan kontekstual antar kata dalam suatu kalimat, terutama pada teks yang panjang dan kompleks [13]. Studi yang dilakukan pada tahun 2019 oleh Sun et al. [14]. Mengkonfirmasi bahwa BERT berhasil mencapai performa terbaik (state-of-the-art) dalam berbagai tugas klasifikasi teks, termasuk analisis sentimen. Keunggulan ini menjadi lebih signifikan pada dataset yang mengandung bahasa informal, seperti komentar di media sosial yang sering kali mencakup singkatan, emotikon, serta penggunaan bahasa slang.

Dalam konteks penelitian ini, komentar di platform YouTube sering kali menggunakan bahasa yang tidak terstruktur dengan beragam ekspresi emosional. Penelitian yang dilakukan

pada tahun 2022 oleh Al Jannah dan Hermawan [15]. Menunjukkan bahwa BERT mampu menangani teks berbahasa Indonesia dengan baik, termasuk teks dengan bahasa informal dan variasi ejaan yang tidak baku. Hal ini disebabkan oleh pelatihan BERT pada korpus yang luas, yang mencakup berbagai sumber teks termasuk media sosial, sehingga model ini lebih adaptif terhadap keragaman bahasa dalam komentar YouTube. Selain itu, efektivitas BERT dalam analisis sentimen teks berbahasa Indonesia telah dibuktikan dalam berbagai penelitian. BERT dirancang sebagai model pembelajaran dua arah yang mampu memahami hubungan kontekstual antar kata dalam teks dengan memanfaatkan arsitektur *Transformer*. Dalam model BERT, setiap kalimat yang diberikan sebagai masukan diubah menjadi representasi vektor untuk setiap kata. Representasi ini dikenal sebagai token, yang berfungsi sebagai unit dasar dalam pemrosesan teks oleh model.

Pada implementasinya, BERT mengubah setiap kalimat menjadi representasi vektor untuk setiap kata atau *token*. Mekanisme *token embeddings* dalam BERT memiliki keunikan dibandingkan Transformer lainnya karena menggunakan *Positional Embeddings* untuk merepresentasikan posisi setiap *token* dalam urutan teks. Dalam urutan tersebut, token pertama selalu berupa token klasifikasi khusus [CLS], diikuti oleh token-token teks pertama, token pemisah [SEP], dan token-token teks kedua hingga seluruh urutan selesai diproses[16]. Selanjutnya, token-token tersebut akan melalui proses *encoding* yang terdiri dari dua tahap utama. Tahap pertama adalah kombinasi antara *multi-head attention* dan *add-norm*, sedangkan tahap kedua adalah kombinasi *feed-forward* dan *add-norm*. Komponen *multi-head attention* bertugas membantu *encoder* menentukan kata-kata yang relevan dengan tetap mempertimbangkan konteks global dari seluruh kata dalam masukan yang diberikan [17]. BERT terdiri dari enam lapisan Transformer yang dibangun di atas arsitektur *encoder* dan *decoder*. Proses pemrosesan dalam BERT diawali dengan representasi kata dalam bentuk embedding yang dihasilkan oleh lapisan embedding. Setiap lapisan Transformer kemudian menghitung *multi-headed attention* berdasarkan representasi kata dari lapisan sebelumnya untuk menghasilkan representasi perantara baru. Semua representasi perantara tersebut memiliki dimensi yang seragam. Pada model BERT yang terdiri dari 12 lapisan, setiap token akan memiliki 12 representasi perantara[12].

2.4.1 Indobenchmark/indobert-base-p1

Dalam penelitian ini, model yang digunakan untuk analisis sentimen adalah *IndoBERT-base-p1*, yang merupakan salah satu varian dari model *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan khusus untuk bahasa Indonesia. Pemilihan

model ini didasarkan pada beberapa pertimbangan metodologis dan teknis yang relevan dengan karakteristik data penelitian. Sejumlah penelitian terdahulu menunjukkan bahwa *IndoBERT-base-p2* memiliki keunggulan dibandingkan *IndoBERT-base-p1*, terutama dalam tugas pemrosesan bahasa alami yang melibatkan teks dengan struktur formal [18].Hal ini disebabkan oleh cakupan korpus pelatihan *IndoBERT-base-p2* yang lebih luas dan beragam dibandingkan *IndoBERT-base-p1*. Namun, meskipun *IndoBERT-base-p2* menunjukkan kinerja lebih baik dalam pemrosesan teks formal, tidak serta-merta model ini lebih unggul dalam semua tugas klasifikasi teks, terutama yang berkaitan dengan bahasa percakapan dan media sosial. Pemilihan model juga ini didasarkan pada studi yang dilakukan pada tahun 2023 oleh Widiansyah et al. [19]. Yang menemukan bahwa IndoBERT-base-p1 lebih efektif dalam memproses teks berbahasa Indonesia yang bersifat informal, seperti komentar di media sosial. IndoBERT-base-p1 telah dilatih pada korpus besar yang mencakup berbagai sumber teks, termasuk media sosial, sehingga lebih adaptif terhadap variasi bahasa yang digunakan dalam komentar YouTube.

Dalam konteks penelitian ini, komentar YouTube yang digunakan sebagai dataset memiliki karakteristik yang berbeda dibandingkan teks akademik atau hukum. Komentar di platform YouTube sering kali mengandung bahasa informal, slang, singkatan, serta variasi ejaan yang tidak baku. Model *IndoBERT-base-p1* telah dilatih menggunakan korpus bahasa Indonesia yang mencakup teks dari berbagai sumber yang lebih dekat dengan bahasa sehari-hari, sehingga lebih mampu menangkap pola bahasa informal yang umum digunakan dalam media sosial [20].Selain itu, pemilihan *IndoBERT-base-p1* juga mempertimbangkan efisiensi komputasi dan kestabilan pelatihan model. Berdasarkan beberapa studi, *IndoBERT-base-p2* memiliki jumlah parameter yang lebih besar dan membutuhkan sumber daya komputasi yang lebih tinggi untuk pelatihan dan inferensi dibandingkan *IndoBERT-base-p1*.Dalam penelitian ini, proses fine-tuning dan eksperimen model dilakukan menggunakan platform *Google Colaboratory* dengan GPU terbatas, sehingga penggunaan *IndoBERT-base-p1* memberikan efisiensi yang lebih baik tanpa mengorbankan akurasi secara signifikan.

Selain dibandingkan dengan *IndoBERT-base-p2*, model ini juga memiliki keunggulan dibandingkan varian lainnya, seperti *IndoBERT-lite* atau *IndoBERT-uncased. IndoBERT-lite* memiliki arsitektur yang lebih ringan dengan jumlah parameter lebih sedikit, sehingga kurang optimal dalam menangkap hubungan kontekstual dalam teks yang kompleks. Sementara itu, *IndoBERT-uncased* tidak mempertahankan informasi huruf kapital, yang dalam beberapa kasus dapat memengaruhi interpretasi makna dalam bahasa Indonesia, terutama dalam analisis sentimen yang bergantung pada konteks kata [18].

Dengan mempertimbangkan karakteristik data, efisiensi komputasi, serta tujuan penelitian ini, penggunaan *IndoBERT-base-p1* dinilai lebih sesuai dibandingkan varian IndoBERT lainnya. Model ini mampu menangani teks dengan bahasa informal, memiliki kompleksitas yang cukup untuk tugas klasifikasi sentimen, serta memungkinkan pelatihan yang lebih efisien dalam lingkungan komputasi yang tersedia. Oleh karena itu, meskipun *IndoBERT-base-p2* memiliki keunggulan dalam beberapa aspek, *IndoBERT-base-p1* dipilih karena lebih sesuai dengan kebutuhan dan keterbatasan penelitian ini.