BAB 2
LANDASAN TEORI

2.1 Text Classification

Text classification merupakan suatu teknik dalam ranah *machine learning* yang bertujuan untuk secara otomatis mengkategorikan teks yang tidak terstruktur ke dalam sejumlah kategori yang telah ditentukan [10]. Dengan memanfaatkan teknik-teknik *machine learning*, *text classification* mampu mengenali pola dan ciri-ciri tertentu dalam teks, memungkinkan pengelompokan yang efisien berdasarkan sifat, ciri, atau fungsi yang telah di identifikasikan [11]. Implementasi *text classification* tidak hanya membantu dalam penyaringan data, namun juga meningkatkan kemampuan sistem untuk memberikan informasi yang relevan dengan cepat melalui pengelompokan otomatis.

Text classification merupakan teknologi yang semakin berkembang dan memiliki aplikasi luas di berbagai industri. Dengan adanya *text classification* membantu meningkatkan efisiensi operasional dalam berbagai sektor [12]. Sebagai contoh dalam dunia bisnis, perusahaan dapat menggunakan *text classification* untuk menganalisis umpan balik pelanggan secara real-time, mengidentifikasi tren atau masalah yang muncul, dan merespon dengan cepat untuk meningkatkan kepuasan pelanggan [13]. Dalam industri kesehatan, *text classification* dapat digunakan untuk mengkategorikan catatan medis, membantu dokter dan peneliti menemukan informasi penting dengan lebih mudah, dan meningkatkan kualitas perawatan pasien [14].

Terdapat juga sektor pendidikan, algoritma ini dapat digunakan untuk mengorganisir makalah penelitian, artikel jurnal, dan materi pembelajaran, sehingga memudahkan mahasiswa dan akademisi untuk mengakses sumber daya yang mereka butuhkan [15]. Kemampuan *text classification* untuk mengolah dan mengkategorikan informasi secara otomatis menawarkan berbagai manfaat yang signifikan di berbagai sektor. Dengan terus berkembangnya teknologi *machine learning* dan peningkatan dalam teknik-teknik analisis data, potensi aplikasi *text classification* diperkirakan akan semakin luas dan canggih, membantu berbagai industri dalam mengoptimalkan proses bisnis dan meningkatkan interaksi manusia dengan sistem digital [14] [15].

2.2 Algoritma BERT

Algoritma BERT, atau *Bidirectional Encoder Representations from Transformers*, merupakan algoritma dalam pemrosesan bahasa alami atau *natural language processing* (NLP) yang dikembangkan oleh Google [16]. Inovasi utama BERT terletak pada kemampuannya untuk memahami kata-kata dalam konteksnya secara *bidirectional*, mengatasi keterbatasan model sebelumnya yang hanya memperhatikan konteks dari satu arah [17]. Melalui *pre-training* pada tugas-tugas besar, seperti prediksi kata yang hilang atau memahami hubungan antara dua kalimat, BERT memperoleh pemahaman mendalam tentang struktur bahasa.

Terdapat aspek penting yang terdapat dalam algoritma BERT dimana aspek tersebut adalah mekanisme *attention* [18]. Mekanisme ini memungkinkannya mempertimbangkan pentingnya berbagai kata dalam sebuah kalimat saat memprosesnya yang membuat BERT dapat menangkap hubungan dan ketergantungan yang rumit dalam konteks kalimat, sehingga meningkatkan pemahamannya tentang bahasa. Hal yang dilakukan oleh algoritma BERT secara signifikan meningkatkan kemampuan model untuk memahami dan menghasilkan representasi teks yang bermakna, sehingga menghasilkan performa yang lebih unggul dalam melakukan berbagai tugas NLP [19].

Algoritma BERT menggunakan arsitektur *transformer*, memungkinkannya untuk memproses urutan data dengan efektif [20]. Kelebihan BERT juga terletak pada fleksibilitasnya, dapat diadaptasi untuk berbagai tugas NLP tanpa perlu mengubah arsitektur inti. Selain itu, BERT menghasilkan *representasi* kata yang bersifat kontekstual dan dapat di "*fine-tune*" untuk tugas spesifik setelah *pre-training*. Dengan kombinasi fitur-fitur tersebut, BERT telah menjadi model unggulan dalam berbagai aplikasi NLP, meningkatkan kinerja pada tugas-tugas seperti klasifikasi teks, pengenalan entitas berbasis teks, dan lainnya [21]. Berikut persamaan dari Algoritma BERT [9] [19].

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

Keterangan:

1. Q, K, dan V adalah matriks yang mewakili vektor Query, Key, dan Value, secara berturut-turut.
2. $\frac{QK^T}{\sqrt{dk}}$ menggambarkan hasil perkalian dot Q dan K^T yang kemudian dibagi dengan akar kuadrat dari dimensi vektor Key ($\sqrt{dk}$)

3. Fungsi softmax ialah mengonversi hasil perkalian dot yang telah di-scaling menjadi bobot *attention*, menghasilkan distribusi probabilitas yang menjumlah menjadi 1.
4. softmax $\left(\frac{QK^T}{\sqrt{dk}}\right) \cdot V$ menggambarkan kombinasi linear dari vektor Value (V) menggunakan bobot *attention* yang dihasilkan.

2.3 Natural Language Processing

Natural language processing (NLP) merupakan teknik yang menggabungkan bidang linguistik, ilmu komputer, dan kecerdasan buatan untuk memungkinkan komputer memahami, menafsirkan, dan menghasilkan bahasa manusia [22]. Dengan memanfaatkan berbagai metode dan algoritma canggih, NLP berupaya untuk menguraikan struktur dan makna dari teks atau ucapan yang dihasilkan oleh manusia. Tujuan utamanya adalah untuk menciptakan interaksi antara manusia dan komputer yang lebih alami, layaknya komunikasi antar manusia pada umumnya [23]. Melalui penerapan NLP, mesin dapat mengidentifikasi dan menganalisis pola bahasa, memahami konteks, serta memberikan respon yang relevan dan akurat, sehingga interaksi menjadi lebih efisien dan efektif .

Penerapan NLP mencakup berbagai aplikasi praktis yang semakin memperkaya pengalaman pengguna dalam kehidupan sehari-hari. Asisten virtual seperti Siri dan Alexa, sistem pencarian informasi, analisis sentimen dalam media sosial, dan pengklasifikasian dokumen adalah beberapa contoh penggunaan NLP [24]. Teknologi ini memungkinkan sistem untuk beradaptasi dengan kebutuhan pengguna secara lebih personal dan responsif. Sebagai hasilnya, teknologi dapat memberikan layanan yang lebih tepat guna, baik dalam konteks pekerjaan, pendidikan, hiburan, maupun kehidupan sehari-hari, sehingga interaksi dengan teknologi menjadi lebih intuitif dan bermanfaat [25].

2.4 Pemodelan Algoritma BERT

Pemodelan BERT merujuk pada penggunaan arsitektur BERT (*Bidirectional Encoder Representations from Transformers*) yang dirancang untuk memahami konteks dalam teks dengan cara yang sangat efisien dan mendalam [26]. BERT menggunakan arsitektur *transformer* yang memungkinkan model untuk memperhatikan dan memahami hubungan antara kata-kata dalam teks dari kedua arah yang dikenal sebagai pendekatan *bidirectional*. Terdapat beberapa tahapan

pada perancangan model BERT yaitu, tahap *pre-training*, *fine-tuning*, *input representation*, *model architecture*, *training* dan *evaluation* [27].

2.4.1 Pre-Training

Pre-training pada model BERT adalah tahap awal di mana, model dilatih pada tugas-tugas umum menggunakan corpus teks besar untuk membangun pemahaman mendalam tentang bahasa alami. Dalam proses ini, BERT dilatih dengan dua tugas utama: *Masked Language Model* (MLM) dan *Next Sentence Prediction* (NSP). Pada tugas MLM, sejumlah kata dalam teks input diacak dan digantikan dengan token khusus [MASK], dan model ditugaskan untuk memprediksi kata asli yang tersembunyi berdasarkan konteks yang tersedia dari kedua arah [28]. Tugas ini memungkinkan BERT memahami makna kata-kata dalam berbagai konteks. Pada tugas NSP, model menerima pasangan kalimat dan harus memprediksi apakah kalimat kedua secara logis mengikuti kalimat pertama dalam corpus asli, yang membantu model memahami hubungan antar kalimat. *Pre-training* dilakukan dengan dataset yang sangat besar dan beragam, sehingga model dapat menangkap berbagai nuansa dan struktur bahasa. Setelah tahap *pre-training* selesai, BERT memiliki representasi kontekstual yang kaya yang bisa digunakan untuk berbagai tugas NLP melalui tahap *fine-tuning* dengan dataset yang lebih spesifik dan lebih kecil [29]. Proses *pre-training* ini sangat penting karena memberikan dasar yang kuat bagi BERT untuk melakukan berbagai tugas pemrosesan bahasa alami dengan akurasi tinggi [28].

2.4.2 Fine-Tuning

Setelah tahap *pre-training*, model BERT akan dilanjutkan ke tahap *fine-tuning*, di mana model disesuaikan untuk tugas-tugas spesifik dengan menggunakan dataset yang lebih kecil namun relevan dengan tugas tersebut [29]. *Fine-tuning* melibatkan mengambil model BERT yang telah dilatih secara umum dan mengadaptasinya untuk aplikasi tertentu seperti klasifikasi teks, dan analisis sentimen.

Pada tahap ini, parameter-parameter model BERT yang sudah dipelajari selama *pre-training* digunakan sebagai titik awal. Dataset spesifik untuk tugas tersebut digunakan untuk melatih ulang model dengan tujuan mengoptimalkan kinerja pada tugas tersebut [30]. Proses *fine-tuning* melibatkan penyesuaian akhir

melalui teknik optimisasi seperti *Adam* untuk meminimalkan *loss* pada dataset spesifik. Model diberi input yang sudah dikodekan (*token embeddings*, *segment embeddings*, dan *position embeddings*). *Loss* dihitung berdasarkan seberapa baik prediksi model sesuai dengan label sebenarnya dalam dataset [30]. Dengan *fine-tuning*, BERT dapat diadaptasi untuk berbagai aplikasi NLP dengan tingkat akurasi dan efisiensi yang tinggi, membuatnya sangat fleksibel dan berguna untuk beragam tugas pemrosesan bahasa.

2.4.3 Input Representation

Input Representation pada model BERT bertujuan untuk membuat model memahami dan memproses teks secara efektif. Input representation di BERT terdiri dari tiga komponen utama: *token embeddings, segment embeddings,* dan *position embeddings* [30] [31]. Setiap komponen ini berkontribusi dalam menangkap informasi yang diperlukan untuk memahami konteks dalam teks. *Token embeddings* bertujuan untuk mengubah setiap kata atau token dalam teks menjadi representasi vektor yang dapat diproses oleh model. *Segment embeddings* digunakan untuk membedakan antara kalimat pertama dan kalimat kedua dalam tugas *Next Sentence Prediction* (NSP). Dan untuk *position embeddings* bertujuan untuk melakukan pengkodean terhadap posisi setiap token dalam urutan teks agar model dapat memahami urutan dan struktur dari sebuah kalimat [31].

Ketiga komponen ini digabungkan dengan menjumlahkan vektor-vektor mereka untuk setiap token dalam teks, menghasilkan representasi input yang kaya dan kontekstual. Representasi ini kemudian diproses oleh lapisan-lapisan *encoder* dalam arsitektur Transformer BERT, memungkinkan model untuk menangkap makna dan hubungan antara kata-kata secara mendalam dan bidirectional. Dengan representasi input yang kompleks dan terintegrasi ini, BERT mampu menangani berbagai tugas NLP dengan performa yang sangat baik [30].

2.4.4 Model Architecture

Arsitektur model pada algoritma BERT menggunakan struktur *Transformer* yang terdiri dari beberapa lapisan *encoder*. Setiap lapisan *encoder* dalam BERT terdiri dari dua komponen utama: *multi-head self-attention mechanism* dan *feed-forward neural network*. BERT memanfaatkan mekanisme *self-attention* untuk memungkinkan setiap token dalam input teks untuk fokus pada token-token lain di

seluruh urutan teks, baik ke kiri maupun ke kanan, sehingga menangkap konteks secara *bidirectional* [32]. Mekanisme *multi-head self-attention* memungkinkan model untuk mempertimbangkan berbagai aspek dari hubungan antar token secara simultan dan meningkatkan kemampuan model untuk memahami nuansa dan makna yang kompleks dalam teks. Kemudian hasilnya diproses oleh *feed-forward neural network* yang sepenuhnya terhubung, yang kemudian diterapkan pada setiap posisi token secara independen namun menggunakan parameter yang sama [32]. Arsitektur *transformer* yang digunakan oleh BERT memungkinkan pemrosesan teks secara paralel, Hal ini membuat BERT lebih efisien dalam hal komputasi dan lebih efektif dalam menangkap hubungan kontekstual dalam teks.

2.4.5 Training

Pada tahap *training* algoritma BERT, model dilatih menggunakan data dalam jumlah besar untuk mempelajari representasi bahasa yang mendalam dan kontekstual. Tahap *training* merujuk pada proses di mana model BERT disesuaikan dengan dataset spesifik untuk tugas tertentu setelah melalui tahap *fine-tuning* dan *pre-training*. Pada tahap ini, model yang telah di *pre-trained* akan diadaptasi kembali untuk mengoptimalkan kinerjanya pada tugas yang lebih spesifik atau domain tertentu [9] [19]. Tahap *training* ini penting karena memungkinkan model BERT untuk menghasilkan parameter yang disesuaikan dengan tugas atau domain spesifik, sehingga meningkatkan kinerja dan akurasi model dalam menyelesaikan tugas tertentu.

2.4.6 Evaluation

Evaluasi dilakukan untuk mengukur kinerja model setelah *training*. Pada tahap evaluasi, model diuji menggunakan dataset yang tidak dilibatkan dalam *training* untuk memastikan generalisasi dan keakuratan prediksi pada data yang belum pernah dilihat sebelumnya. Metode evaluasi dapat meliputi perhitungan metrik seperti akurasi, nilai *precision*, *recall*, dan *f1 score*, tergantung pada jenis tugas yang dihadapi [33]. Evaluasi ini penting untuk memastikan bahwa model tidak hanya menghafal *data training*, tetapi juga mampu melakukan prediksi yang akurat pada data baru. Dengan proses *training* dan evaluasi yang cermat, BERT dapat diterapkan pada berbagai aplikasi pemrosesan bahasa alami dengan kinerja yang sangat baik.

2.5 United Nations (UN) Sustainable Development Goals (SDG)

United Nations Sustainable Development Goals (UN SDG) merupakan sebuah agenda global yang memiliki tujuan utama untuk menangani berbagai masalah signifikan di dunia. Terdapat 17 tujuan dalam UN SDG yang mencakup beragam isu seperti kemiskinan, kelaparan, pendidikan, kesehatan, kesetaraan gender, air bersih, energi terjangkau, pekerjaan layak, dan perlindungan lingkungan [6]. Diperkenalkan pada tahun 2015, agenda ini diterima dan diakui oleh seluruh negara anggota Perserikatan Bangsa-Bangsa (PBB) serta berbagai pihak pemangku kepentingan, termasuk pemerintah, sektor swasta, dan organisasi masyarakat sipil.

Melalui pendekatan yang holistik dan berkelanjutan, UN SDG bertujuan untuk menciptakan perubahan positif dalam berbagai aspek kehidupan, mendorong pembangunan yang inklusif, serta mengatasi tantangan-tantangan global seperti perubahan iklim dan ketidaksetaraan [34]. Dengan menerapkan kerjasama lintas sektor dan lintas batas, diharapkan UN SDG dapat memberikan solusi berkelanjutan dan membawa dunia menuju masa depan yang lebih baik bagi seluruh penduduk bumi.