BAB 2
LANDASAN TEORI

2.1 Ujaran Kebencian

2.1.1 Definisi Ujaran Kebencian

Ujaran kebencian (*hate speech*) adalah bentuk komunikasi yang menyerang, merendahkan, atau mendiskriminasi individu maupun kelompok berdasarkan karakteristik tertentu seperti ras, agama, etnis, jenis kelamin, orientasi seksual, disabilitas, atau asal kebangsaan [14, 15]. Menurut Mahendra et al. [6], ujaran kebencian dapat mencakup penghinaan, ancaman, serta ajakan kekerasan terhadap kelompok tertentu. Menurut Kamus Besar Bahasa Indonesia (KBBI), istilah "ujaran kebencian" didefinisikan sebagai "ujaran yang menyerukan kebencian terhadap orang atau kelompok tertentu" [16].

Di Indonesia, ujaran kebencian dijelaskan dalam Surat Edaran Kapolri No. SE/6/X/2015 sebagai segala bentuk tindakan yang memuat penghinaan, penistaan, perbuatan tidak menyenangkan, provokasi, penghasutan, dan penyebaran berita bohong yang ditujukan terhadap individu atau kelompok berdasarkan suku, agama, aliran kepercayaan, ras, antargolongan (SARA), serta orientasi seksual, disabilitas fisik dan/atau mental [17].

Undang-Undang No. 40 Tahun 2008 tentang Penghapusan Diskriminasi Ras dan Etnis juga mengakui bentuk ujaran kebencian sebagai pernyataan yang mendiskreditkan atau menyebarkan kebencian terhadap individu atau kelompok berdasarkan ras atau etnis tertentu [18].

2.1.2 Jenis-Jenis Ujaran Kebencian

Ujaran kebencian dapat muncul dalam berbagai bentuk, baik secara eksplisit maupun implisit, dan dapat diklasifikasikan ke dalam beberapa kategori, antara lain [19].:

1. Ujaran Kebencian Rasial: Mengandung hinaan atau diskriminasi berdasarkan ras atau etnis tertentu.

2. Ujaran Kebencian Berbasis Agama: Menyerang kepercayaan, ajaran, atau simbol-simbol agama tertentu.

3. Ujaran Kebencian terhadap Gender atau Seksualitas: Menargetkan individu berdasarkan jenis kelamin, identitas gender, atau orientasi seksual, termasuk misogini, homofobia, dan transfobia.

4. Ujaran Kebencian Berbasis Nasionalisme atau Asal Negara: Merendahkan kelompok berdasarkan asal negara atau kebangsaan.

5. Ujaran Kebencian Terselubung: Menggunakan bahasa yang lebih halus namun tetap menyampaikan kebencian atau ajakan diskriminasi secara tidak langsung

2.1.3 Dampak Ujaran Kebencian

Dampak dari ujaran kebencian bersifat multidimensional, mencakup aspek psikologis, sosial, politik, dan hukum:

1. Dampak Psikologis: Individu yang menjadi sasaran dapat mengalami tekanan psikologis, kehilangan harga diri, dan trauma berkepanjangan [20].

2. Dampak Sosial: Terjadi keretakan hubungan antar kelompok masyarakat, meningkatnya prasangka, dan melemahnya rasa persatuan [19].

3. Dampak Politik: Ujaran kebencian sering dimanfaatkan untuk polarisasi politik dan delegitimasi lawan politik, terutama di masa pemilu [15].

4. Dampak Hukum: Di Indonesia, ujaran kebencian telah diatur dalam beberapa regulasi seperti Undang-Undang Informasi dan Transaksi Elektronik (UU ITE), meskipun implementasinya masih menjadi perdebatan karena potensi benturan dengan kebebasan berekspresi [6].

2.2 Media Sosial

Media sosial merupakan sebuah platform digital yang memungkinkan penggunanya untuk membuat, berbagi, dan saling bertukar informasi, ide, serta konten melalui jaringan virtual. Menurut Kaplan dan Haenlein [21], media sosial adalah “sebuah kelompok aplikasi berbasis internet yang dibangun di atas dasar ideologis dan teknologi dari Web 2.0, yang memungkinkan penciptaan dan pertukaran konten yang dibuat oleh pengguna.”

2.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) bekerja melalui serangkaian proses yang memungkinkan komputer memahami, menganalisis, dan menghasilkan bahasa alami secara efektif. Proses ini merupakan gabungan dari pendekatan linguistik, statistik, serta teknik pembelajaran mesin yang saling melengkapi untuk mengekstrak informasi dari teks atau ucapan dalam bahasa manusia [22].

2.4 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah model bahasa berbasis arsitektur Transformer yang dikembangkan oleh Google AI Language pada tahun 2018 [1]. Model ini menggunakan pendekatan kontekstual dua arah (*bidirectional*) untuk memahami teks, yang memungkinkan BERT mempertimbangkan konteks dari kedua arah (kiri dan kanan) dalam satu kalimat secara simultan [1].

2.4.1 Arsitektur BERT

BERT dibangun sepenuhnya dengan arsitektur Transformer, khususnya bagian encoder yang terdiri dari lapisan *multi-head self-attention* dan *feed-forward neural network* [1]. Tidak seperti model tradisional seperti RNN atau LSTM yang memproses kata secara sekuensial, BERT memproses semua kata secara paralel menggunakan mekanisme self-attention [1].

BERT memiliki dua versi utama:

1. BERT-Base: terdiri dari 12 lapisan encoder, 768 ukuran vektor tersembunyi, dan 12 kepala atensi. Total parameter: sekitar 110 juta.
2. BERT-Large: terdiri dari 24 lapisan encoder, 1024 ukuran vektor tersembunyi, dan 16 kepala atensi. Total parameter: sekitar 340 juta [1].

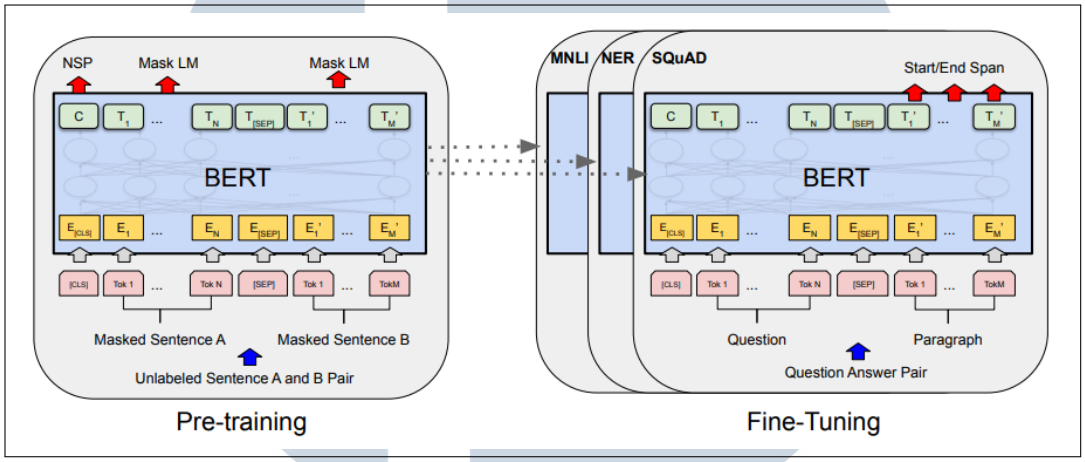
2.4.2 Pre-training Tasks

BERT dilatih secara tidak diawasi (*unsupervised*) menggunakan dua tugas utama, yaitu:

1. Masked Language Modeling (MLM): Sebagian token dalam kalimat disembunyikan (`[MASK]`) dan model dilatih untuk memprediksi token tersebut

berdasarkan konteks dua arah. Ini membuat BERT mampu memahami relasi antar kata secara mendalam [1].

2. Next Sentence Prediction (NSP): Model diberikan dua kalimat dan harus memprediksi apakah kalimat kedua merupakan kelanjutan dari kalimat pertama. NSP membantu model memahami hubungan antar kalimat [1].



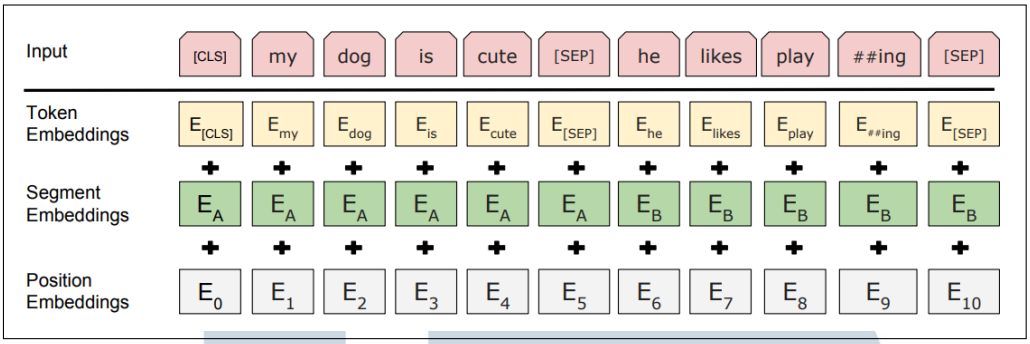
Gambar 2.1. Gambaran Pre-training BERT dan Fine-Tuning BERT [1]

2.4.3 Tokenisasi dan Representasi Input

BERT menggunakan tokenisasi WordPiece, yang memecah kata menjadi sub-kata untuk menangani kata yang tidak dikenal (*out-of-vocabulary*) [1]. Sebelum dimasukkan ke dalam model, setiap input akan direpresentasikan dalam bentuk kombinasi dari tiga jenis embedding, yaitu:

- Token Embeddings: Representasi vektor dari token hasil tokenisasi. Misalnya, kata *playing* dapat dipecah menjadi *play* dan *##ing*.
- Segment Embeddings (Token Type IDs): Digunakan untuk membedakan dua kalimat yang dijadikan pasangan dalam tugas NSP, yaitu segmen A dan segmen B.
- Position Embeddings: Memberikan informasi urutan posisi token dalam kalimat [1].

Ketiga jenis embedding tersebut dijumlahkan secara elemen-wise untuk membentuk representasi akhir setiap token. Ilustrasi dari proses ini ditampilkan pada Gambar 2.2.



Gambar 2.2. Representasi input pada BERT, terdiri dari token embeddings, segment embeddings, dan position embeddings [1].

2.4.4 Fine-tuning BERT

Setelah proses pelatihan awal (*pre-training*), model BERT dapat disesuaikan untuk berbagai tugas pemrosesan bahasa alami melalui tahapan *fine-tuning*. Dalam tahap ini, arsitektur dasar BERT tidak diubah, melainkan ditambahkan satu layer klasifikasi sederhana di atas token khusus [CLS], kemudian seluruh parameter model dilatih ulang secara end-to-end pada dataset spesifik [1].

Dalam penelitian awal oleh Devlin [1], *fine-tuning* BERT dilakukan untuk berbagai tugas seperti klasifikasi teks (contohnya Multi-Genre Natural Language Inference atau MNLI), penjawaban pertanyaan (Stanford Question Answering Dataset atau SQuAD), dan pengenalan entitas bernama (Named Entity Recognition atau NER). Pada tugas klasifikasi teks seperti MNLI, model memprediksi hubungan logis antara dua kalimat. Pada tugas SQuAD, model menjawab pertanyaan berdasarkan konteks paragraf, sedangkan pada tugas NER, model menandai token dalam teks yang merupakan nama entitas seperti orang, lokasi, atau organisasi.

2.4.5 Keunggulan BERT

BERT menawarkan beberapa keunggulan dibandingkan model NLP sebelumnya:

1. Pemahaman kontekstual dua arah yang lebih dalam (*deep bidirectional context*) [1].
2. Kemampuan transfer learning yang kuat untuk berbagai tugas NLP hanya dengan sedikit fine-tuning [1].

3. Performa tinggi pada berbagai benchmark seperti GLUE, SQuAD, dan SWAG [1].

2.5 IndoBERT

IndoBERT adalah model BERT (Bidirectional Encoder Representations from Transformers) yang dilatih secara khusus pada korpus besar berbahasa Indonesia. Model ini dikembangkan oleh IndoNLU (2020) untuk mengatasi tantangan dalam pemrosesan bahasa alami di Indonesia, seperti struktur gramatikal yang unik dan ragam bentuk kata. IndoBERT menggunakan arsitektur BERT-Base dengan 12 lapisan transformer dan 110 juta parameter, serta dilatih pada lebih dari 200 juta kata dari berbagai sumber termasuk berita, media sosial, dan Wikipedia Indonesia [23].

2.6 IndoBERTweet

IndoBERTweet adalah model transformer berbasis BERTweet yang diadaptasi untuk bahasa Indonesia dan dilatih secara khusus pada data dari Twitter. Model ini dirancang untuk menangani karakteristik unik dari cuitan Twitter seperti singkatan, tagar, emoji, dan tata bahasa informal [13].

2.7 Pra-pemrosesan Teks

Pra-pemrosesan teks merupakan tahap penting dalam pemrosesan bahasa alami (Natural Language Processing/NLP). Tahapan ini bertujuan untuk membersihkan dan menyusun teks agar sesuai dengan kebutuhan model pembelajaran mesin [24].

Adapun tahapan-tahapan pra-pemrosesan yang umum digunakan adalah sebagai berikut:

1. Lowercasing, yaitu mengubah seluruh huruf menjadi huruf kecil [25].
2. Penghapusan karakter khusus, termasuk URL, simbol, dan tanda baca yang tidak relevan.
3. Penghapusan stopword, yaitu kata-kata umum yang tidak memiliki makna penting secara kontekstual.

4. Stemming, yakni mengubah kata ke bentuk dasar menggunakan algoritma seperti Nazief-Adriani melalui library Sastrawi [26].

5. Tokenisasi, yaitu memecah teks menjadi unit terkecil (token). Pada penelitian ini, tokenisasi dilakukan menggunakan tokenizer dari IndoBERT dan IndoBERTweet [23].

2.8 Evaluation Metrics

Dalam penelitian ini, untuk mengukur performa model digunakan beberapa metrik evaluasi yang umum pada tugas klasifikasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score* [27].

Accuracy mengukur proporsi prediksi yang benar terhadap total jumlah data dan dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Precision merupakan proporsi dari prediksi positif yang benar dibandingkan seluruh prediksi positif yang dibuat oleh model, dengan rumus sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

Recall, atau sensitivitas, adalah rasio antara prediksi positif yang benar terhadap seluruh data positif sebenarnya. Rumus Recall ditunjukkan pada persamaan berikut:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

Sementara itu, *F1-score* merupakan rata-rata harmonik dari Precision dan Recall yang digunakan untuk memberikan keseimbangan antara keduanya. Rumus F1-score ditunjukkan pada persamaan berikut:

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

Dimana:

1. TP (*True Positive*) adalah jumlah data positif yang diprediksi sebagai positif,
2. TN (*True Negative*) adalah jumlah data negatif yang diprediksi sebagai negatif,
3. FP (*False Positive*) adalah jumlah data negatif yang diprediksi sebagai positif,
4. FN (*False Negative*) adalah jumlah data positif yang diprediksi sebagai negatif.

Metrik ini penting untuk mengevaluasi kualitas model, terutama pada kasus *imbalanced dataset* seperti deteksi ujaran kebencian [28].