BAB 2
LANDASAN TEORI

Pada bab ini dijelaskan teori yang digunakan, model dan penjelasan penelitian terdahulu.

2.1 Seksisme

Seksisme merupakan bentuk diskriminasi berbasis gender yang masih banyak dijumpai dalam ruang diskusi daring. Dalam penelitian yang dilakukan oleh Janah terkait pelecehan seksual, seksisme, dan bystander, dijelaskan bahwa seksisme muncul akibat ketidaksetaraan gender dan sering kali dilestarikan oleh norma sosial tradisional yang menempatkan satu gender sebagai superior atas yang lain [17]. Seksisme biasanya ditandai oleh sikap, keyakinan, dan ekspresi yang merendahkan atau melemahkan individu berdasarkan jenis kelamin [18].

Konsep ambivalent sexism yang diperkenalkan oleh Glick dan Fiske membagi seksisme menjadi dua dimensi utama, yaitu *hostile sexism* (sikap negatif dan permusuhan terbuka terhadap perempuan) dan *benevolent sexism* (sikap yang tampaknya positif namun tetap memperkuat ketimpangan gender melalui pengukuhan peran gender tradisional) [19]. Kedua bentuk ini memungkinkan seksisme hadir tidak hanya dalam bentuk agresi eksplisit, tetapi juga dalam bentuk perlakuan merendahkan yang terselubung dalam narasi kepedulian atau pelindungan [19].

Di media sosial, ekspresi seksis sering muncul dalam komentar pengguna, meme, atau video pendek. Mohammadi *et al.* menunjukkan bahwa seksisme digital banyak diwujudkan melalui bahasa yang menonjolkan objektifikasi, stereotip, misogini, maupun kekerasan seksual [18]. Fenomena ini diperkuat oleh norma budaya patriarkis dan sifat anonim media daring, yang kerap membuat pelaku merasa lepas dari tanggung jawab atas ujarannya [17].

Untuk menangkap kompleksitas bentuk seksisme di ruang digital, kompetisi EXIST 2024 mengelompokkan ekspresi seksis ke dalam lima kategori utama: (1) *ideological and inequality*, yakni narasi yang mendiskreditkan feminisme atau menyangkal ketimpangan gender; (2) *stereotyping and dominance*, yaitu asumsi bahwa perempuan tidak layak menempati posisi atau peran tertentu; (3) *objectification*, yakni reduksi perempuan menjadi objek seksual; (4) *sexual*

violence, berupa komentar atau ajakan bernada pelecehan; dan (5) *misogyny and non-sexual violence*, yang mencakup ujaran kebencian terhadap perempuan [20]. Klasifikasi ini menjadi pijakan penting dalam pengembangan model NLP untuk deteksi otomatis konten seksis secara komprehensif.

Dalam konteks media sosial, seksisme menjadi semakin kompleks karena sifat ruang digital yang cepat, luas, dan cenderung tanpa batas geografis maupun regulasi ketat. Perempuan sering kali menjadi kelompok paling terdampak oleh penyebaran pandangan seksis secara daring, mengingat adanya fitur-fitur seperti komentar anonim, algoritma berbasis tren, serta dinamika popularitas yang mendorong keterlibatan emosional tinggi [21]. Bentuk seksisme yang tersebar di platform seperti TikTok atau Twitter tidak selalu bersifat vulgar atau eksplisit, melainkan bisa hadir dalam bentuk yang lebih halus dan terselubung, seperti candaan seksis, stereotip gender, atau pengucilan sosial dalam diskusi publik. Misalnya, komentar yang menyiratkan bahwa perempuan kurang rasional dalam berdiskusi, atau bahwa nilai perempuan hanya terletak pada penampilan fisiknya, adalah bentuk seksisme yang kerap dinormalisasi [22].

Untuk memahami dinamika ini secara lebih konseptual, Glick dan Fiske (1996) mengembangkan teori *ambivalent sexism*, yang menyatakan bahwa seksisme terdiri dari dua dimensi utama: *hostile sexism* dan *benevolent sexism*.

1. ***Hostile sexism*** merujuk pada sikap permusuhan yang eksplisit terhadap perempuan. Ini mencakup keyakinan bahwa perempuan berusaha mengendalikan laki-laki melalui manipulasi emosional atau bahwa mereka tidak layak menempati posisi otoritas. Bentuk ini sering muncul dalam ujaran kebencian, ancaman, dan komentar yang merendahkan kapasitas intelektual atau moral perempuan.

2. ***Benevolent sexism***, sebaliknya, bersifat lebih halus dan tampak positif. Ia tercermin dalam perlakuan ”pengayoman” terhadap perempuan yang diasumsikan lemah, emosional, atau memerlukan perlindungan dari laki-laki. Contoh dari benevolent sexism adalah anggapan bahwa perempuan lebih cocok di ranah domestik, atau bahwa perempuan harus dijaga kesuciannya oleh laki-laki. Meskipun tampak menghargai, sikap ini sesungguhnya menempatkan perempuan pada posisi subordinat yang tidak setara secara struktural.

Kedua bentuk seksisme ini saling melengkapi dan memperkuat struktur patriarki dalam masyarakat. Studi empiris menunjukkan bahwa baik *hostile*

maupun *benevolent sexism* berkorelasi dengan tingkat penerimaan terhadap mitos pemerkosaan, kekerasan dalam hubungan intim, dan pembatasan partisipasi perempuan dalam ruang publik [23]. Dengan demikian, pemahaman terhadap seksisme harus melampaui narasi kekerasan verbal eksplisit, dan mencakup dimensi-dimensi simbolik dan struktural yang lebih tersembunyi namun tetap merugikan perempuan secara sistemik.

Berdasarkan kerangka kerja yang dikembangkan dalam kompetisi *EXIST 2024* (salah satu benchmark internasional dalam tugas identifikasi seksisme daring), seksisme digital diklasifikasikan ke dalam lima kategori utama yang merepresentasikan berbagai bentuk manifestasi ujaran seksis di media sosial [24]. Kategori-kategori tersebut tidak hanya mencerminkan bentuk kekerasan verbal yang eksplisit, tetapi juga bentuk diskriminasi gender yang bersifat simbolik dan struktural, sebagai berikut:

1. ***Ideological and Inequality***
   Kategori ini mencakup pernyataan yang mendiskreditkan gerakan feminisme, menyangkal adanya ketimpangan gender, atau bahkan membalikkan narasi ketimpangan dengan mengklaim bahwa laki-laki adalah korban penindasan berbasis gender. Misalnya, komentar seperti "Feminisme hanya ingin menjatuhkan laki-laki" atau "Sekarang laki-laki yang tertindas" adalah bentuk ujaran yang termasuk dalam kategori ini. Sikap ini sering kali bertujuan untuk melemahkan legitimasi perjuangan kesetaraan gender dan memperkuat status quo patriarki.

2. ***Stereotyping and Dominance***
   Merupakan bentuk seksisme yang mengandung stereotip terhadap perempuan, seperti anggapan bahwa perempuan secara alami emosional, tidak rasional, atau hanya cocok mengurus rumah tangga. Selain itu, kategori ini juga mencakup dominasi simbolik, yaitu pernyataan bahwa laki-laki lebih kompeten atau layak memegang peran penting dibandingkan perempuan. Contohnya adalah komentar seperti "Perempuan nggak cocok kerja di bidang teknologi" atau "Kalau jadi bos, perempuan pasti baperan."

3. ***Objectification***
   Dalam kategori ini, perempuan diperlakukan atau digambarkan semata-mata sebagai objek seksual, tanpa memperhatikan nilai, pikiran, atau kepribadian mereka. Ini termasuk komentar yang menilai perempuan hanya berdasarkan

fisiknya, seperti “Paha cewe itu yang bikin semangat kerja” atau “Kalau nggak cantik, nggak layak masuk kamera.” Reduksi perempuan menjadi objek seksual bukan hanya merendahkan martabat individu, tetapi juga memperkuat pemakluman terhadap pelecehan seksual.

4. ***Sexual Violence***
   Merujuk pada pesan-pesan yang mengandung unsur pelecehan, ajakan seksual yang tidak diinginkan, hingga ancaman kekerasan seksual secara verbal. Komentar seperti “Mau nggak aku ajarin enak di ranjang?” atau “Bisa kok aku perkosa kamu cuma lewat chat” termasuk dalam kategori ini. Meskipun dilakukan secara daring, dampaknya terhadap korban tetap nyata, termasuk trauma psikologis dan rasa tidak aman.

5. ***Misogyny and Non-Sexual Violence***
   Kategori ini mencakup ujaran kebencian dan kekerasan non-seksual yang ditujukan kepada perempuan secara umum, termasuk hasutan untuk menyakiti, merendahkan martabat, atau menolak keberadaan perempuan dalam ruang publik. Contohnya seperti “Perempuan cuma bikin ribet, mending dihapus aja dari dunia” atau “Kalau perempuan ngelawan, pantas dipukul.” Bentuk ini merupakan eskalasi dari seksisme menjadi misogini, yaitu kebencian yang mendalam terhadap perempuan sebagai kelompok sosial.

Klasifikasi ini menunjukkan bahwa seksisme dalam ruang digital tidak bersifat homogen, melainkan mencerminkan beragam bentuk penindasan gender yang saling terkait. Dengan memahami kelima kategori ini, proses deteksi otomatis terhadap komentar seksis seperti yang dilakukan dalam penelitian ini dapat lebih akurat dan kontekstual dalam mengidentifikasi pola ujaran bermasalah yang tersembunyi dalam interaksi daring.

2.2 Validasi Cohen’s Kappa

Dalam penelitian yang melibatkan pelabelan data secara manual, pengukuran tingkat kesepakatan antar-penilai (inter-rater agreement atau IRR) menjadi sangat krusial untuk menjamin reliabilitas dan objektivitas data yang dihasilkan [25]. IRR merupakan metode yang berkontribusi pada peningkatan transparansi dan konsistensi analisis data, terutama dalam bagaimana kode dan

konstruksi dikembangkan dari data mentah İRR menyediakan indikator krusial yang mengungkapkan sejauh mana penilai independen melihat artefak yang sama dan mencapai kesimpulan yang sama.

Salah satu statistik yang umum digunakan untuk data nominal (kategorikal) adalah koefisien Kappa Cohen (κ) [26]. Metrik ini lebih kuat dibandingkan persentase kesepakatan sederhana karena secara khusus mengoreksi kesepakatan yang terjadi secara kebetulan atau acak. Hal ini berarti nilai Kappa memberikan estimasi kesepakatan yang lebih jujur dan objektif, dengan mengecualikan kesepakatan yang kebetulan semata.

Nilai Cohen's Kappa dihitung berdasarkan proporsi kesepakatan yang diamati (P_o) dan proporsi kesepakatan yang diharapkan terjadi secara kebetulan (P_e) [26]. Rumus perhitungan Cohen's Kappa adalah sebagai berikut:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.1}$$

Di mana:

1. P_o merepresentasikan probabilitas kesepakatan yang diamati antara penilai.
2. P_e adalah probabilitas kesepakatan yang diharapkan terjadi secara acak atau kebetulan.

Nilai Kappa berkisar antara -1.0 hingga $+1.0$. Nilai 1 menunjukkan kesepakatan sempurna, nilai 0 menunjukkan kesepakatan yang hanya terjadi karena kebetulan, dan nilai negatif menunjukkan kesepakatan yang lebih buruk dari kebetulan. Interpretasi nilai Cohen's Kappa seringkali mengikuti pedoman standar. Menurut pedoman ini, nilai Kappa dapat diinterpretasikan sebagai berikut:

1. Below 0.00 : *Poor agreement*
2. 0.00 – 0.20 : *Slight agreement*
3. 0.21 – 0.40 : *Fair agreement*
4. 0.41 – 0.60 : *Moderate agreement*
5. 0.61 – 0.80 : *Substantial agreement*
6. 0.81 – 1.00 : *Almost perfect agreement*

Dalam konteks pelabelan data, nilai Kappa yang tinggi (misalnya di atas 0.60) sangat diinginkan karena menunjukkan bahwa label yang diberikan oleh penilai bersifat konsisten dan dapat diandalkan. Ini merupakan prasyarat untuk menghasilkan kualitas data yang baik dalam pelatihan model klasifikasi. Proses rekonsiliasi IRR, yang memaksa perbedaan interpretasi untuk dihadapi dan mendorong pembangunan konsensus, memiliki potensi untuk meningkatkan teori yang muncul dari analisis data.

2.3 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) merupakan disiplin dalam kecerdasan buatan yang memungkinkan mesin untuk memahami, menghasilkan, dan menafsirkan bahasa manusia secara otomatis. Sejak diperkenalkannya arsitektur Transformer, mekanisme self-attention memfasilitasi pemodelan hubungan antar kata dalam kalimat secara paralel dan efisien, sehingga mampu mengatasi keterbatasan model sekuensial seperti RNN dan LSTM dalam memahami konteks panjang dan kompleks [**?**]. Arsitektur ini menjadi fondasi bagi berbagai model NLP modern seperti BERT, GPT, dan RoBERTa yang telah terbukti memberikan performa tinggi dalam berbagai tugas bahasa alami seperti klasifikasi teks, ekstraksi entitas, dan deteksi ujaran kebencian [**?**].

Kemajuan tersebut turut mendorong pengembangan model bahasa lokal seperti IndoBERT, yang diadaptasi khusus untuk Bahasa Indonesia dengan menerapkan prinsip arsitektur Transformer dan proses pretraining pada korpus berbahasa lokal [9]. Oleh karena itu, Transformer menjadi basis yang kokoh dalam pengembangan NLP modern, terutama untuk mendukung tugas-tugas kompleks seperti pendeteksian komentar seksis di media sosial [27].

2.4 *Text Classification*

Klasifikasi teks (*text classification*) merupakan salah satu permasalahan fundamental dalam bidang *Natural Language Processing* (NLP) yang bertujuan menetapkan satu atau lebih label pada unit teks seperti kalimat, paragraf, atau dokumen. Tugas ini menjadi semakin penting seiring meningkatnya volume data tekstual secara eksponensial dari berbagai sumber, seperti media sosial, layanan pelanggan, email, dan berita daring [28]. Aplikasi klasifikasi teks mencakup berbagai bidang, antara lain analisis sentimen, deteksi spam, kategorisasi berita,

question answering, *natural language inference*, hingga moderasi konten [29].

Secara umum, pendekatan klasifikasi teks terbagi menjadi dua kategori utama, yakni berbasis aturan (*rule-based*) dan berbasis pembelajaran mesin (*machine learning*). Perkembangan terbaru menunjukkan bahwa pendekatan berbasis *deep learning*, khususnya model *transformer*, telah melampaui metode konvensional dalam berbagai tugas klasifikasi karena kemampuannya dalam menangkap konteks dan relasi semantik kompleks antar kata [28]. Selain itu, kebutuhan akan sistem klasifikasi yang akurat dan minim bias telah mendorong penelitian untuk merancang arsitektur klasifikasi teks yang lebih efisien dan adaptif terhadap domain aplikasi tertentu [29]. Dalam konteks penelitian ini, klasifikasi teks digunakan untuk mengidentifikasi komentar yang mengandung seksisme dalam konten media sosial berbahasa Indonesia, sebagai tahap awal dalam pengembangan sistem deteksi otomatis yang andal.

2.5 Model BERT dan Representasi Input

2.5.1 Arsitektur dan Tujuan Pre-training

BERT (Bidirectional Encoder Representations from Transformers) adalah model bahasa berbasis arsitektur Transformer encoder yang dilatih secara dua arah, artinya model memperhitungkan konteks dari sisi kiri dan kanan suatu token secara simultan [30]. Model ini diperkenalkan oleh Devlin et al. pada tahun 2019 dan telah menjadi fondasi dari banyak tugas NLP modern.

Pre-training BERT dilakukan dengan dua tujuan utama:

1. **Masked Language Modeling (MLM)**: Sebagian token dalam input diganti dengan token `[MASK]`, dan model dilatih untuk menebak token yang hilang berdasarkan konteksnya.

2. **Next Sentence Prediction (NSP)**: Model dilatih untuk memprediksi apakah dua kalimat yang diberikan saling berurutan secara logis dalam korpus asli.

Proses pre-training ini memungkinkan BERT memahami struktur bahasa secara umum sebelum diterapkan pada tugas spesifik melalui *fine-tuning*.

2.5.2 Struktur Input dan Token Khusus

Sebelum suatu teks diproses oleh BERT, teks tersebut harus dikonversi ke dalam bentuk token-token yang sesuai dengan format input standar BERT. Format

tersebut terdiri dari:

- `[CLS]`: Token klasifikasi yang selalu muncul di awal input. Representasi akhir dari token ini digunakan untuk tugas klasifikasi.
- Token-token dari kalimat pertama.
- `[SEP]`: Digunakan untuk menandai akhir dari kalimat pertama dan juga pemisah antar dua kalimat.
- Token-token dari kalimat kedua (jika ada).
- `[SEP]`: Ditambahkan di akhir untuk menutup input.

Sebagai contoh, dua kalimat seperti "saya suka kopi" dan "saya tidak suka teh" akan direpresentasikan sebagai:

```
[CLS] saya suka kopi [SEP] saya tidak suka teh [SEP]
```

2.5.3 Tokenisasi Subword dengan WordPiece

BERT tidak menggunakan tokenisasi berbasis kata utuh, melainkan menggunakan metode subword bernama *WordPiece* [30]. Algoritma ini bertujuan untuk menangani permasalahan kata-kata langka atau tidak dikenal (*out-of-vocabulary*) dengan memecah kata menjadi unit-unit sub-kata berdasarkan frekuensi kemunculan dalam korpus.

Sebagai contoh:

```
berjualan → ['ber', 'jual', '##an']
```

Tanda `##` menunjukkan bahwa token tersebut merupakan lanjutan dari token sebelumnya.

IndoBERT, sebagai turunan BERT yang dilatih menggunakan korpus Bahasa Indonesia, menggunakan 31.923 token WordPiece yang diperoleh dari korpus berita dan sosial media Bahasa Indonesia [31].

2.5.4 Representasi Embedding Input BERT

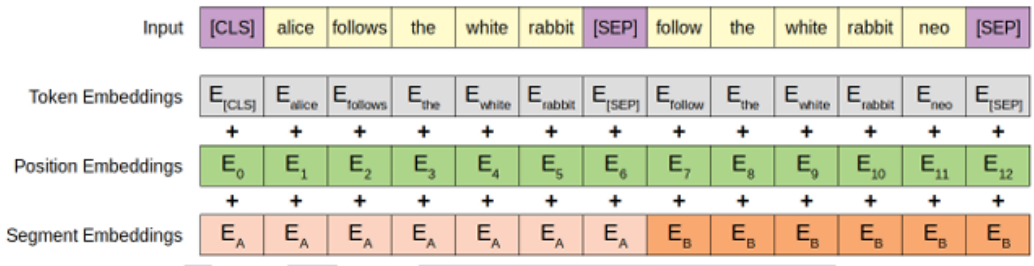
Setiap token yang telah ditokenisasi kemudian dikonversi menjadi vektor numerik atau *embedding*. Dalam BERT, embedding akhir untuk token ke-i, dilambangkan sebagai $\mathbf{E}_i$, merupakan hasil penjumlahan dari tiga komponen:

$$\mathbf{E}_i = \mathbf{T}_i + \mathbf{P}_i + \mathbf{S}_i \tag{2.2}$$

dengan:

- $\mathbf{T}_i$: **Token Embedding**, yaitu representasi dari token berdasarkan kamus WordPiece.
- $\mathbf{P}_i$: **Position Embedding**, yaitu vektor yang menyandikan posisi token dalam urutan input, karena Transformer tidak memiliki mekanisme urutan secara implisit [32].
- $\mathbf{S}_i$: **Segment Embedding**, yaitu vektor yang menyatakan apakah token berasal dari kalimat pertama (segmen A) atau kalimat kedua (segmen B).

2.5.5 Visualisasi Embedding Token



Gambar 2.1. Proses pembentukan embedding dari token input BERT. Diadaptasi dari [30].

Gambar 2.1 menggambarkan tahapan pembentukan vektor input BERT:

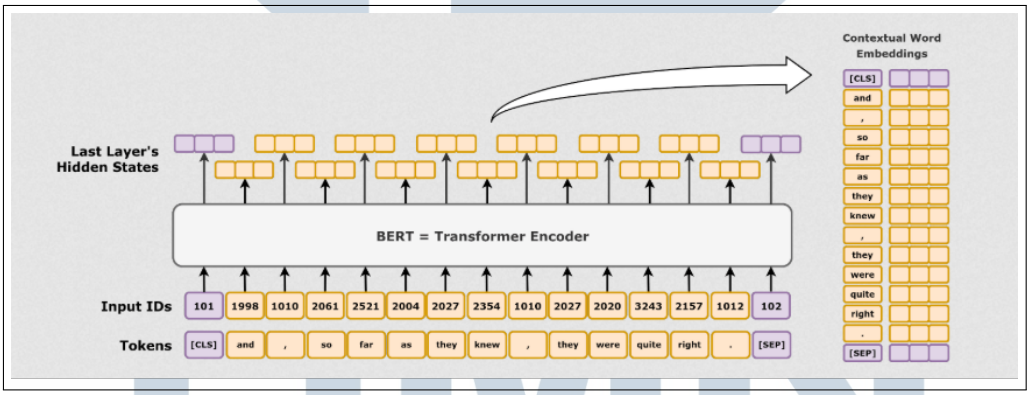
1. Kalimat di-tokenisasi menjadi unit subword, dimulai dengan `[CLS]` dan diakhiri dengan `[SEP]`.
2. Setiap token diberi:
   - Token Embedding ($\mathbf{T}_i$),

- Position Embedding ($\mathbf{P}_i$) berdasarkan posisinya,
- Segment Embedding ($\mathbf{S}_i$) sesuai asal kalimatnya.

3. Ketiga embedding tersebut dijumlahkan untuk membentuk representasi final $\mathbf{E}_i$ dari token ke-i.
4. Seluruh rangkaian vektor $\mathbf{E}_1, \ldots, \mathbf{E}_n$ kemudian menjadi input bagi encoder Transformer.

2.5.6 Dimensi Embedding dan Representasi Kontekstual

Pada model BERT-base, setiap token input dikodekan menjadi vektor berdimensi $\mathbb{R}^{768}$, yang dikenal sebagai *token embedding*. Vektor ini merupakan hasil penjumlahan dari tiga jenis embedding, yaitu token embedding, segment embedding, dan position embedding. Ketiga embedding ini dikombinasikan sebagai representasi awal token sebelum diproses oleh Transformer.



Gambar 2.2. Alur Representasi Kontekstual pada BERT: Dari Token hingga Hidden State . Diadaptasi dari [33]

Sebagaimana ditunjukkan pada Gambar 2.2, token-token yang telah dikonversi ke dalam ID numerik kemudian dimasukkan ke dalam encoder BERT. Di dalam encoder, terdapat 12 lapisan (*layers*) Transformer yang memproses informasi secara bertingkat melalui mekanisme *multi-head self-attention*. Proses ini menghasilkan representasi kontekstual (*contextualized embedding*) dari setiap token, yaitu representasi akhir yang mencerminkan relasi semantik dan sintaktik antar token dalam satu kalimat.

Secara khusus, token `[CLS]` yang ditambahkan di awal input memiliki peran penting dalam tugas klasifikasi. Representasi akhir dari token `[CLS]`, yaitu vektor

$\mathbf{h}_{[CLS]} \in \mathbb{R}^{768}$, dianggap sebagai ringkasan dari keseluruhan konteks kalimat. Vektor ini kemudian diteruskan ke lapisan linear dan fungsi softmax untuk menghasilkan probabilitas kelas (misalnya: kelas 0 atau 1 dalam deteksi seksisme).

2.6 IndoBERT: Adaptasi BERT untuk Bahasa Indonesia

Model BERT asli [30] dilatih menggunakan korpus berbahasa Inggris seperti BooksCorpus dan Wikipedia, sehingga tidak optimal saat diaplikasikan pada bahasa lain seperti Bahasa Indonesia. Hal ini mendorong pengembangan model BERT khusus untuk Bahasa Indonesia, yaitu **IndoBERT**, yang dirilis oleh Koto et al. dalam proyek IndoLEM [31].

IndoBERT bertujuan untuk menyediakan representasi bahasa kontekstual yang lebih sesuai dengan struktur morfologi, sintaksis, dan semantik Bahasa Indonesia. Model ini menggunakan arsitektur yang sama dengan BERT-base, namun dilatih dari awal menggunakan korpus dalam Bahasa Indonesia.

2.7 IndoBERT: Adaptasi BERT untuk Bahasa Indonesia

IndoBERT adalah model BERT yang dikembangkan secara khusus untuk Bahasa Indonesia. Berbeda dari BERT asli yang dilatih menggunakan korpus berbahasa Inggris, IndoBERT dilatih dari awal (*from scratch*) menggunakan korpus besar berbahasa Indonesia, termasuk data dari Wikipedia, berita online (seperti Kompas, Tempo, Detik), serta media sosial [31].

Model ini menggunakan arsitektur BERT-base yang sama dengan versi aslinya [30], dengan 12 layer transformer, 768 dimensi hidden size, dan 12 head attention. Namun, perbedaan utama terletak pada korpus pre-training dan vocabulary WordPiece yang dikustomisasi untuk Bahasa Indonesia, sebanyak 31.923 token.

IndoBERT juga mengikuti strategi pre-training BERT, yaitu Masked Language Modeling (MLM) dan Next Sentence Prediction (NSP). Setelah pre-training, model dapat difine-tune untuk berbagai tugas NLP seperti klasifikasi teks, Named Entity Recognition (NER), atau deteksi komentar seksis sebagaimana dilakukan dalam penelitian ini.

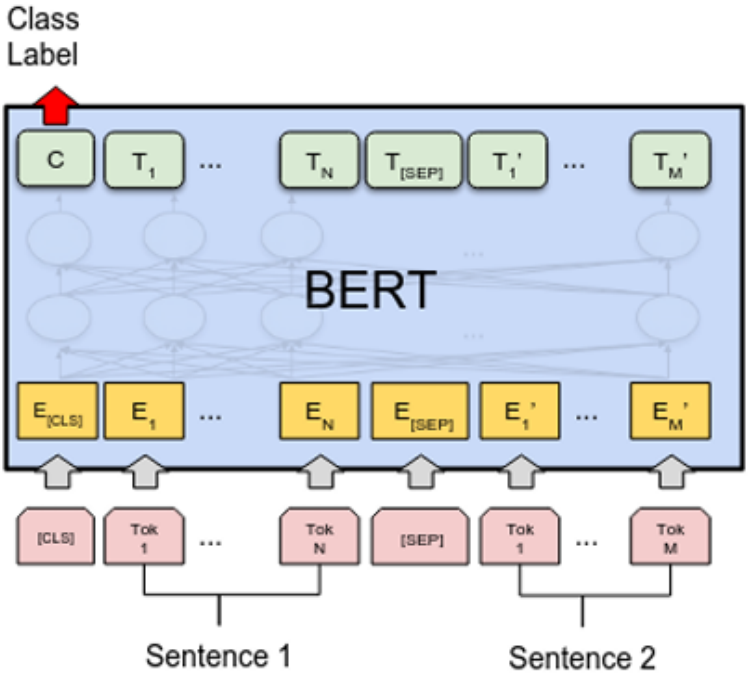
Keunggulan IndoBERT terletak pada kemampuannya memahami struktur morfologi khas Bahasa Indonesia, seperti imbuhan (awalan dan akhiran), bentuk pasif, serta variasi bahasa formal dan informal, yang kurang ditangani oleh model

multilingual.

2.8 Fine-tuning IndoBERT untuk Klasifikasi Komentar Seksis

Dalam tugas klasifikasi komentar seksis, model IndoBERT digunakan untuk mengubah input teks menjadi representasi vektor yang kaya konteks, lalu memetakan representasi tersebut ke dalam label kelas (seksis atau tidak seksis). Proses ini terdiri dari tiga tahap utama yaitu tokenisasi dan embedding, pemrosesan dengan encoder BERT, serta klasifikasi menggunakan linear layer dan fungsi aktivasi softmax. Penjelasan masing-masing tahap diperjelas melalui tiga gambar berikut.

2.8.1 Representasi Umum Arsitektur BERT



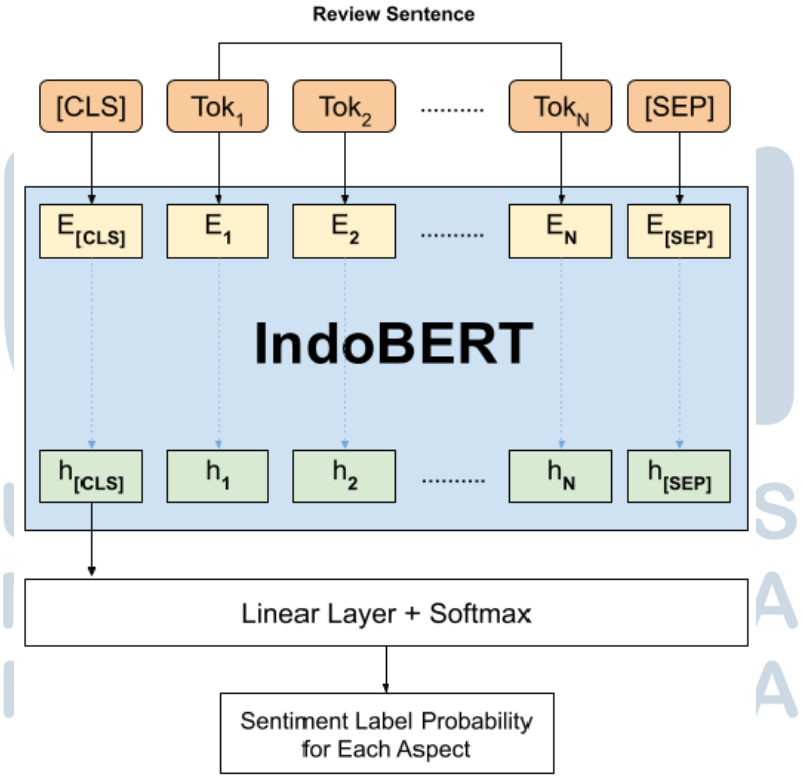
Gambar 2.3. Representasi Arsitektur BERT
. Diadaptasi dari [30].

Gambar 2.3 memperlihatkan struktur dasar BERT ketika menerima input kalimat. Token-token dalam setiap kalimat ditambahkan dengan token khusus seperti [CLS] di awal dan [SEP] sebagai pemisah antar kalimat. Setiap token diubah menjadi representasi vektor embedding, kemudian diproses oleh lapisan-lapisan

encoder dalam BERT. Token [CLS] digunakan sebagai representasi global dari seluruh kalimat, dan hasil akhirnya digunakan dalam proses klasifikasi. Mekanisme pembobotan konteks dilakukan melalui self-attention yang mengatur seberapa besar perhatian diberikan dari satu token ke token lain, berdasarkan relasi semantiknya.

Gambar 2.1 menjelaskan bagaimana setiap token input diubah menjadi vektor embedding melalui penjumlahan tiga komponen: token embedding, position embedding, dan segment embedding. Token embedding merupakan representasi vektor dari sub-kata yang telah ditokenisasi, position embedding menyandikan informasi urutan token dalam kalimat, dan segment embedding membedakan apakah token berasal dari kalimat pertama atau kedua. Proses penjumlahan ketiganya menghasilkan vektor akhir untuk setiap token, yang kemudian digunakan sebagai input bagi lapisan transformer dalam IndoBERT. Representasi ini memungkinkan model menangkap fitur linguistik dan semantik dari teks secara otomatis tanpa rekayasa fitur manual.

2.8.2 Fine-tuning IndoBERT untuk Klasifikasi Kalimat Tunggal



Gambar 2.4. Fine-tuning IndoBERT untuk Klasifikasi Kalimat Tunggal . Diadaptasi dari [34].

Gambar 2.4 menunjukkan proses fine-tuning IndoBERT untuk klasifikasi satu kalimat, seperti pada tugas deteksi komentar seksis. Input berupa kalimat dikonversi menjadi token dan diberikan token [CLS] dan [SEP]. Setelah melalui embedding, token-token tersebut diproses oleh encoder IndoBERT yang menghasilkan representasi kontekstual untuk masing-masing token. Vektor keluaran dari token [CLS], yaitu $h_{[\text{CLS}]}$, dianggap sebagai representasi seluruh kalimat dan digunakan sebagai input ke layer klasifikasi.

Layer klasifikasi terdiri dari linear layer yang menghasilkan nilai logit untuk masing-masing kelas, lalu diproses menggunakan fungsi aktivasi softmax:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.3}$$

dengan z_i adalah logit kelas ke-i dan K adalah jumlah kelas (2 dalam kasus ini). Hasil dari softmax berupa probabilitas untuk masing-masing kelas, dan kelas dengan probabilitas tertinggi dipilih sebagai prediksi akhir. Sebagai contoh, jika probabilitas kelas 1 (seksis) adalah 0.87 dan kelas 0 (tidak seksis) adalah 0.13, maka model akan mengklasifikasikan input tersebut sebagai seksis. Dengan pendekatan ini, IndoBERT mampu melakukan klasifikasi secara end-to-end tanpa memerlukan fitur buatan tangan dan menghasilkan prediksi berbasis pemahaman konteks yang mendalam.

2.9 K-Fold Cross-Validation

Cross-validation merupakan teknik evaluasi statistik yang bertujuan untuk mengurangi bias dalam estimasi performa model yang dapat muncul akibat pembagian data pelatihan dan pengujian yang statis. Teknik ini memberikan evaluasi yang lebih stabil dan representatif terhadap kemampuan generalisasi model pada data yang tidak terlihat.

Salah satu metode cross-validation yang paling umum digunakan adalah k-fold cross-validation, di mana dataset dibagi menjadi k subset (fold) yang berukuran seimbang. Model kemudian dilatih dan diuji sebanyak k kali. Pada setiap iterasi ke-i, $k-1$ subset digunakan sebagai data latih, sementara 1 subset sisanya digunakan sebagai data uji [35].

Setelah seluruh iterasi selesai, performa model dari setiap fold dirata-ratakan untuk memperoleh estimasi performa keseluruhan. Proses ini membantu mengurangi varian hasil evaluasi akibat pemilihan subset tertentu sebagai data uji

[36].

Perhitungan rata-rata skor evaluasi dari k-fold cross-validation dinyatakan dalam Persamaan 2.4:

$$\text{CV}_{\text{mean}} = \frac{1}{k} \sum_{i=1}^{k} \text{Score}_i \tag{2.4}$$

Keterangan:

- CV_{mean}: Rata-rata skor evaluasi model dari seluruh fold.
- k: Jumlah fold atau bagian pembagi dataset.
- Score_i: Nilai metrik evaluasi (misalnya akurasi, F1-score) pada fold ke-i.

Untuk menggambarkan proses pelatihan dan pengujian model secara formal pada tiap iterasi cross-validation, digunakan rumus berikut:

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathscr{L}(f^{(i)}, D_i) \tag{2.5}$$

Keterangan:

- $\text{CV}_{(k)}$: Estimasi performa rata-rata model dari seluruh fold.
- $\mathscr{L}$: Fungsi loss atau metrik evaluasi yang digunakan (misalnya cross-entropy, akurasi, F1-score, dan sebagainya).
- $f^{(i)}$: Model hasil pelatihan pada data selain D_i (yakni $D \setminus D_i$).
- D_i: Subset ke-i yang digunakan sebagai data uji pada iterasi ke-i.

Penerapan metode ini sangat penting dalam tugas klasifikasi teks seperti deteksi ujaran seksis, mengingat data yang digunakan sering kali tidak seimbang (*imbalanced*) dan mengandung noise kontekstual. Sebagaimana dijelaskan oleh Zhang dan Yang [37], penggunaan cross-validation secara konsisten mampu mengurangi varian hasil evaluasi dan membantu dalam pemilihan model yang lebih stabil dan andal terhadap data baru yang tidak terlihat selama pelatihan.

2.10 *Hyperparameter Tuning*

Hyperparameter tuning adalah proses eksploratif untuk menemukan kombinasi nilai optimal dari parameter-parameter eksternal yang tidak dipelajari

langsung oleh model, seperti *learning rate*, *dropout rate*, jumlah epoch, dan ukuran batch. Nilai-nilai ini sangat memengaruhi kemampuan model dalam belajar dan generalisasi. Pemilihan nilai hyperparameter yang tidak tepat dapat menyebabkan masalah *underfitting* atau *overfitting* pada model.

Terdapat beberapa pendekatan umum dalam tuning hyperparameter, seperti *Grid Search*, *Random Search*, dan *Bayesian Optimization*. Menurut Kobayashi et al., pendekatan *Bayesian Optimization* terbukti lebih efisien daripada *Grid Search* karena mampu menemukan kombinasi optimal dengan jumlah iterasi yang lebih sedikit [38].

Untuk menghindari potensi bias dalam proses pemilihan model saat melakukan *hyperparameter tuning*, diterapkan praktik *nested cross-validation*. Dalam pendekatan ini, data dibagi ke dalam dua tingkat validasi: *outer loop* digunakan untuk mengevaluasi performa model akhir, sementara *inner loop* digunakan untuk tuning hyperparameter. Skema ini direkomendasikan terutama untuk model yang kompleks seperti fine-tuned IndoBERT karena mampu memberikan estimasi performa yang lebih objektif [39].

Dengan memisahkan proses validasi dan tuning secara sistematis, penelitian ini memastikan bahwa model deteksi komentar seksis pada TikTok dapat mencapai performa optimal sekaligus menjaga generalisasi dan akurasi pada data baru.

2.11 *Evaluation Metrix*

Untuk mengevaluasi performa model klasifikasi, digunakan sejumlah metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Metrik-metrik ini tidak hanya mengukur seberapa banyak prediksi yang benar, tetapi juga mempertimbangkan jenis kesalahan yang dilakukan model terutama penting dalam konteks dataset yang tidak seimbang. Keempat metrik ini dihitung berdasarkan nilai-nilai dalam *confusion matrix*.

Confusion matrix adalah sebuah tabel yang menggambarkan perbandingan antara label sebenarnya dengan label yang diprediksi oleh model. Untuk klasifikasi biner, confusion matrix terdiri dari empat komponen utama, seperti yang ditunjukkan pada Tabel 2.1.

Tabel 2.1. Contoh confusion matrix untuk klasifikasi biner

	Prediksi Positif	**Prediksi Negatif**
Aktual Positif	True Positive (TP)	False Negative (FN)
Aktual Negatif	False Positive (FP)	True Negative (TN)

Penjelasan masing-masing komponen sebagai berikut:

- **True Positive (TP)**: Kasus di mana model memprediksi kelas positif dan label sebenarnya juga positif.
- **False Positive (FP)**: Model memprediksi kelas positif padahal sebenarnya negatif (juga disebut *Type I Error*).
- **False Negative (FN)**: Model memprediksi kelas negatif padahal sebenarnya positif (*Type II Error*).
- **True Negative (TN)**: Model memprediksi kelas negatif dan label sebenarnya juga negatif.

Dari nilai-nilai ini, metrik evaluasi dihitung menggunakan rumus berikut:

2.11.1 Accuracy

Accuracy mengukur proporsi prediksi yang benar terhadap seluruh prediksi yang dilakukan oleh model. Metrik ini umum digunakan pada dataset dengan distribusi kelas yang seimbang. Namun, akurasi dapat menjadi metrik yang menyesatkan ketika digunakan pada dataset yang memiliki distribusi kelas tidak merata.

Rumus 2.6 digunakan untuk menghitung nilai *Accuracy*, yaitu proporsi prediksi yang benar terhadap keseluruhan jumlah prediksi. Metode ini sangat umum digunakan dalam evaluasi performa model klasifikasi biner.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

Keterangan:

1. TP (*True Positive*): Jumlah komentar seksis yang diklasifikasikan secara benar sebagai seksis.

2. TN (*True Negative*): Jumlah komentar tidak seksis yang diklasifikasikan secara benar sebagai tidak seksis.

3. FP (*False Positive*): Jumlah komentar tidak seksis yang salah diklasifikasikan sebagai seksis.

4. FN (*False Negative*): Jumlah komentar seksis yang salah diklasifikasikan sebagai tidak seksis.

Semakin tinggi nilai *Accuracy*, maka semakin baik performa model dalam mengklasifikasikan komentar dengan benar. Namun, dalam kasus ketidakseimbangan data (*imbalanced data*), metrik ini sebaiknya dilengkapi dengan *precision*, *recall*, dan *F1-score* untuk memperoleh evaluasi performa yang lebih menyeluruh.

2.11.2 Precision

Precision mengukur ketepatan model dalam memprediksi kelas positif, yaitu seberapa besar bagian dari seluruh prediksi positif yang benar-benar positif.

Rumus 2.7 menunjukkan cara perhitungan *Precision*, yaitu rasio antara jumlah prediksi positif yang benar terhadap seluruh prediksi positif yang dihasilkan oleh model. Metrik ini penting terutama ketika kesalahan dalam klasifikasi positif (False Positive) berdampak besar, seperti dalam konteks deteksi komentar seksis.

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

Dengan definisi masing-masing variabel sebagai berikut:

1. TP (*True Positive*): Jumlah komentar yang benar-benar seksis dan diklasifikasikan sebagai seksis oleh model.

2. FP (*False Positive*): Jumlah komentar yang sebenarnya tidak seksis namun salah diklasifikasikan sebagai seksis oleh model.

Precision mengukur ketepatan model dalam memprediksi kelas positif (seksis). Semakin tinggi nilai precision, semakin sedikit komentar tidak seksis yang salah dikenali sebagai seksis.

2.11.3 Recall

Recall (juga dikenal sebagai sensitivitas) mengukur seberapa baik model dapat menangkap semua instance yang benar-benar termasuk dalam kelas positif.

Rumus 2.8 menunjukkan cara perhitungan *Recall*, yaitu rasio antara jumlah prediksi positif yang benar terhadap seluruh data yang seharusnya diklasifikasikan sebagai positif. Metrik ini sangat penting dalam konteks deteksi komentar seksis, karena mengukur kemampuan model dalam menangkap sebanyak mungkin komentar yang benar-benar seksis.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.8}$$

Dengan definisi masing-masing variabel sebagai berikut:

1. TP (*True Positive*): Jumlah komentar yang benar-benar seksis dan berhasil diklasifikasikan sebagai seksis oleh model.
2. FN (*False Negative*): Jumlah komentar yang sebenarnya seksis tetapi gagal dikenali oleh model dan diklasifikasikan sebagai tidak seksis.

Recall mengukur sensitivitas model terhadap kelas positif (seksis). Semakin tinggi nilai recall, semakin baik model dalam mengidentifikasi komentar-komentar yang seharusnya terdeteksi sebagai seksis.

2.11.4 F1-Score

F1-Score merupakan rata-rata harmonik dari precision dan recall. Metrik ini berguna ketika dibutuhkan keseimbangan antara precision dan recall, khususnya pada kasus dengan data tidak seimbang.

Rumus 2.9 menunjukkan cara perhitungan *F1-Score*, yaitu metrik harmonik yang menggabungkan *Precision* dan *Recall*. Metrik ini sangat berguna ketika terdapat ketidakseimbangan data antara kelas positif dan negatif, seperti pada deteksi komentar seksis yang sering kali memiliki jumlah data tidak seimbang.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.9}$$

Dengan definisi masing-masing variabel sebagai berikut:

1. **Precision**: Rasio antara jumlah prediksi positif yang benar terhadap seluruh prediksi positif yang dilakukan oleh model.

2. **Recall**: Rasio antara jumlah prediksi positif yang benar terhadap seluruh data yang seharusnya diklasifikasikan sebagai positif.

F1-Score memberikan keseimbangan antara *Precision* dan *Recall*. Nilai yang tinggi menunjukkan bahwa model tidak hanya tepat dalam mendeteksi kelas positif, tetapi juga sensitif terhadap keberadaan kelas tersebut.

2.12 Penelitian Terkait

Tabel 2.2. Ringkasan Penelitian Terdahulu

Peneliti (Tahun)	**Judul Penelitian**	**Algoritma / Model**	**Preprocessing dan Split Data**	**Hasil Evaluasi**
Kusuma & Chowanda (2023)	Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter	IndoBERTweet + BiLSTM	1. Lowercasing, hapus RT, @user, URL, emoji 2. Normalisasi slang 3. Split: 80% train, 20% test; dari train → 80% train, 20% val 4. Max token: 128	1. Akurasi: 93.7% 2. Epoch: 5 (BiLSTM) 3. Batch Size: 10 4. LR: $1e^{-5}$

Bersambung ke halaman berikutnya

Tabel 2.2 – lanjutan dari halaman sebelumnya

Peneliti (Tahun)	**Judul Penelitian**	**Algoritma / Model**	**Preprocessing dan Split Data**	**Hasil Evaluasi**
Darmawan et al. (2023)	Experiments on IndoBERT Implementation for Detecting Multi-Label Hate Speech with Data Resampling through Synonym Replacement Method	IndoBERT (base)	1. Normalisasi slang dan typo 2. Case folding 3. Hapus URL, user, tanda baca, angka, karakter non-alfabet 4. Stopword removal 5. Stemming (PySastrawi) 6. Split data: 3 variasi (1) 80:10:10, (2) 70:15:15, (3) 60:20:20 7. Stratifikasi kelas untuk imbalanced dataset 8. Data augmentation menggunakan synonym replacement (dua cara: per kata dan semua kata)	1. Akurasi: 88.23% 2. Epoch: 15 3. Batch Size: 32 4. LR: $2e^{-5}$

Bersambung ke halaman berikutnya

Tabel 2.2 – lanjutan dari halaman sebelumnya

Peneliti (Tahun)	**Judul Penelitian**	**Algoritma / Model**	**Preprocessing dan Split Data**	**Hasil Evaluasi**
Wijanarko et al. (2024):	Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts	IndoBERTweet (fine-tuned)	Stratified 10-Fold Cross-Validation	Accuracy 0.89 untuk hate speech

Bersambung ke halaman berikutnya

Tabel 2.2 – lanjutan dari halaman sebelumnya

Peneliti (Tahun)	**Judul Penelitian**	**Algoritma / Model**	**Preprocessing dan Split Data**	**Hasil Evaluasi**
Bremm et al. (2024)	Detecting Sexism in German Online Newspaper Comments with Open-Source Text Embeddings (GerMS-Detect, GermEval2024)	Text Embedding: mE5-large dan GBERT-large-pc	1. Train/val split: 80% train, 20% validation 2. 5-fold cross-validation 3. Annotator-specific training sets 4. Final Training: (1) Menggunakan seluruh data train (5998), dan (2) 10% data untuk early stopping 5. Hyperparameter tuning dengan grid search 6. Penanganan class imbalance: class weights atau oversampling 7. Embeddings tidak dilatih ulang (hanya classifier yang dilatih)	1. Model terbaik: SVM + GBERT-large-pc (peringkat 1) 2. Jensen-Shannon Distance: 0.301 (peringkat 2)

Penelitian yang dilakukan oleh Kusuma dan Chowanda membahas deteksi ujaran kebencian pada media sosial Twitter menggunakan model *IndoBERTweet* yang dikombinasikan dengan arsitektur BiLSTM sebagai lapisan tambahan [7].

Studi ini dilatarbelakangi oleh meningkatnya penyebaran ujaran kebencian di platform daring dan perlunya sistem otomatis untuk membantu mengidentifikasi konten berbahaya secara efisien. Dalam eksperimen mereka, penulis menggunakan dataset publik berlabel yang terdiri dari dua kelas: ujaran kebencian dan non-ujaran kebencian. Model *IndoBERTweet* + *BiLSTM* terbukti memberikan performa terbaik dibandingkan baseline seperti CNN dan *Fine-tuned IndoBERTweet*, dengan capaian akurasi 93.7%, *recall* 92.9%, *precision* 93.8%, dan *F1-score* sebesar 93.3%. Hasil ini menunjukkan bahwa kombinasi representasi token kontekstual dari IndoBERTweet dan kapabilitas sekuensial BiLSTM mampu menangkap struktur bahasa yang kompleks dalam konteks ujaran kebencian secara lebih akurat dibandingkan pendekatan lain. Penelitian ini menjadi bukti efektivitas model berbasis *transformer* yang diperkuat dengan arsitektur *RNN* dalam menangani tugas klasifikasi teks pada Bahasa Indonesia.

Penelitian dalam [8] bertujuan untuk mendeteksi ujaran kebencian multikelas dalam Bahasa Indonesia dengan memanfaatkan model *IndoBERT* dan metode *synonym replacement* untuk mengatasi permasalahan ketidakseimbangan data. Studi ini mengkaji berbagai kombinasi distribusi data, teknik praproses, dan strategi stratifikasi dataset untuk mengoptimalkan performa klasifikasi. Dalam total 12 eksperimen yang dilakukan, model terbaik diperoleh dari kombinasi dataset tanpa praproses dan distribusi stratifikasi sebesar 80% data latih, 10% validasi, dan 10% data uji, menghasilkan akurasi makro sebesar 88,23%. Selain klasifikasi biner HS (Hate Speech), penelitian ini juga menguji akurasi terhadap beberapa label spesifik seperti *HS_Gender* (97,94%), *HS_Physical* (97,49%), dan *HS_Strong* (96,50%), menunjukkan efektivitas model dalam menangani tugas klasifikasi multikelas. Strategi augmentasi berbasis sinonim terbukti mampu meningkatkan representasi label minoritas serta memperbaiki performa klasifikasi secara keseluruhan.

Penelitian dalam [9] mengembangkan sistem klasifikasi ujaran kebencian berbasis NLP dengan fokus pada konteks Indonesia, terutama selama pemilu presiden 2024. Peneliti mengklasifikasikan ujaran kebencian ke dalam beberapa kategori, seperti *profanity or obscenity*, *insult*, *incitement to violence*, *identity attack*, dan *sexual explicit*. Model *IndoBERTweet*, yang dilatih menggunakan dataset IndoToxic2024 dan dievaluasi menggunakan *10-fold cross-validation*, menunjukkan performa superior dibanding model lain seperti GPT-3.5 dan SeaLLM, dengan skor *macro-F1* tertinggi sebesar 0,718. Pada masing-masing tugas klasifikasi, model ini menunjukkan akurasi tertinggi sebesar 96% untuk deteksi

ujaran terkait pemilu, dan nilai *macro-F1* tertinggi sebesar 0,93 pada kategori yang sama. Sementara itu, performa lebih rendah ditemukan pada kategori *incitement to violence* dengan *macro-F1* sebesar 0,53. Sistem ini juga diimplementasikan dalam bentuk dasbor daring yang digunakan oleh lembaga seperti BAWASLU dan AJI untuk memantau ujaran kebencian terhadap kelompok rentan di media sosial.

Penelitian yang dilakukan oleh Bremm, Blaneck, Bornheim, Grieger, dan Bialonski [10] bertujuan untuk mendeteksi komentar seksis dan misoginis dalam media daring berbahasa Jerman menggunakan representasi teks berbasis *open-source text embeddings*. Dataset yang digunakan berasal dari komentar surat kabar daring Austria dan dievaluasi dalam kompetisi GerMS-Detect 2024. Arsitektur model yang digunakan melibatkan tiga jenis *text embedding*, yaitu *GBERT-large-pc*, *mE5-base*, dan *mE5-large*, yang semuanya digunakan tanpa proses fine-tuning tambahan. Representasi vektor dari komentar dimasukkan ke dalam tiga algoritma klasifikasi: *Multilayer Perceptron (MLP)*, *Random Forest Classifier (RFC)*, dan *Support Vector Machine (SVM)*.

Penelitian ini menggunakan teknik *grid search* dengan 5-fold cross-validation untuk melakukan tuning hiperparameter seperti `hidden_layer_sizes`, `n_estimators`, dan `C` pada masing-masing algoritma. Masalah ketidakseimbangan kelas ditangani dengan penyesuaian `class_weight` pada semua model.

Hasil evaluasi menunjukkan bahwa model SVM dengan embedding *mE5-large* memberikan performa terbaik pada tugas klasifikasi biner (Subtask 1) dengan skor macro-F1 sebesar 0,597 dan menempati peringkat ke-4 pada leaderboard Codabench. Pada tugas prediksi distribusi anotasi manusia (Subtask 2), model SVM dengan embedding *GBERT-large-pc* memperoleh rata-rata Jensen-Shannon Distance sebesar 0,301 dan menduduki peringkat ke-2. Penelitian ini menyimpulkan bahwa pendekatan berbasis embedding terbuka dan SVM mampu mereplikasi penilaian manusia dalam mendeteksi seksisme secara efisien serta menunjukkan potensi skalabilitas pada berbagai konteks bahasa dan budaya yang berbeda.

Penelitian yang dilakukan oleh Fudulu *et al.* [11] mengusulkan pendekatan berbasis *ensemble transformer* untuk mendeteksi seksisme daring dalam kompetisi SemEval 2023 Task 10: Explainable Detection of Online Sexism (EDOS). Fokus utama penelitian ini adalah Task A (klasifikasi biner: seksis atau tidak) dan Task B (klasifikasi multi-label penjelas).

Untuk Task A, peneliti menggunakan strategi *ensemble voting* dari tiga model transformer terbaik yang telah diseleksi, yaitu `ernie-2.0-base-en`,

`bert-base-uncased`, dan `microsoft/deberta-v3-base`. Masing-masing model menjalani proses *hyperparameter tuning* menggunakan *grid search* terhadap subset data yang terdiri atas 500 tweet seksis dan 500 non-seksis. Parameter yang disesuaikan meliputi *learning rate* (2×10^{-5}), *batch size* (16–32), *weight decay* (0,1–0,001), dan *max sequence length* (32–64).

Pada tahap pengembangan, model ensemble berhasil memperoleh skor *macro F1* sebesar 0,8403 untuk Task A dan 0,6467 untuk Task B. Sementara itu, pada fase pengujian resmi kompetisi, performa sistem mencapai skor 0,8396 untuk Task A (peringkat ke-33) dan 0,5794 untuk Task B (peringkat ke-51). Penelitian ini menegaskan bahwa pendekatan berbasis *ensemble learning* dan pencarian parameter optimal mampu meningkatkan akurasi dan efisiensi dalam mendeteksi serta mengklasifikasikan teks seksis secara daring [11].

Dari tinjauan penelitian sebelumnya, dapat disimpulkan bahwa model berbasis Transformer seperti *IndoBERT* dan variannya sangat efektif untuk tugas deteksi ujaran kebencian dan seksisme dalam Bahasa Indonesia. Namun, mayoritas penelitian berfokus pada platform Twitter atau menggunakan dataset umum. Penelitian yang secara spesifik meneliti komentar pada platform dinamis seperti TikTok, yang cenderung menggunakan bahasa lebih informal dan kontekstual, masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan menerapkan *IndoBERT* secara khusus untuk mendeteksi seksisme dalam komentar TikTok berbahasa Indonesia.