Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3716~3725
e-ISSN: 2548-6861

Comparison of Online Gambling Promotion Detection Performance Using DistilBERT and DeBERTa Models

Halim Meliana Pratama [1], IGN Lanang Wijayakusuma [2], Ratna Sari Widiastuti [3]
Matematika, Universitas Udayana
melianaprtm016@student.unud.ac.id [1], lanang_wijaya@unud.ac.id [2], ratnasariwidiastuti@unud.ac.id

Article Info

Article history:

Received 2025-09-22
Revised 2025-11-23
Accepted 2025-11-26

Keyword:

DistilBERT,
DeBERTa,
Online Gambling,
Transformer,
Text Classification.

ABSTRACT

Online gambling promotions on social media have become a serious concern in Indonesia, where perpetrators use ambiguous and disguised language to evade detection. This study compares two transformer-based models, DistilBERT and DeBERTa, in detecting such content within Indonesian YouTube comments. Using a balanced dataset of 6,350 comments, both models were fine-tuned with optimized hyperparameters (learning rate 1e-5, batch size 32, 5 epochs) and evaluated through five-fold cross-validation. Results show that DeBERTa achieves superior performance with 99.84% accuracy and perfect recall, while DistilBERT achieves 99.29% accuracy. Error and linguistic analyses indicate that DeBERTa's disentangled attention and Byte-Pair Encoding provide better understanding of non-standard and ambiguous language. Despite requiring higher computational cost, DeBERTa is ideal for high-accuracy applications, whereas DistilBERT remains suitable for real-time and resource-limited environments.

I. PENDAHULUAN

Era digital telah mengubah interaksi sosial masyarakat secara fundamental. Platform media sosial seperti YouTube, Facebook, Instagram, TikTok, dan X kini menjadi bagian dalam kehidupan sehari-hari jutaan pengguna di Indonesia. Data menunjukkan bahwa pada tahun 2024, pengguna media sosial di Indonesia mencapai 191 juta pengguna, dengan YouTube sebagai platform terpopuler yang memiliki 139 juta pengguna, diikuti TikTok (98 juta), Instagram (90 juta), dan X (27 juta) [1]. Pertumbuhan eksponensial ini mencerminkan besarnya potensi media sosial sebagai media informasi dan interaksi, namun di sisi lain, juga membuka celah bagi berbagai bentuk penyalahgunaan konten, khususnya penyebaran promosi judi online.

Fenomena judi online telah menjadi permasalahan serius yang mengancam kesehatan sosial dan ekonomi masyarakat Indonesia. Para pelaku menggunakan strategi penyamaran yang canggih dengan memanfaatkan bahasa ambigu, istilah-istilah terselubung, dan frasa yang tidak langsung untuk menghindari deteksi oleh sistem moderasi konten. Promosi judi online seringkali menyasar segmen pengguna muda yang lebih familiar dengan teknologi digital, sehingga dampak negatifnya dapat meluas ke berbagai aspek kehidupan. Metode moderasi konten secara manual terbukti tidak lagi memadai untuk menangani volume dan kompleksitas konten yang terus berkembang. Keterbatasan sumber daya manusia, subjektivitas dalam penilaian, sertaketidakmampuan dalam memproses data berskala besar secara real-time menjadi hambatan utama dalam upaya deteksi dan pencegahan penyebaran promosi judi online.

Beberapa penelitian telah menunjukkan efektivitas berbagai pendekatan dalam deteksi konten judi online. Penelitian pada tahun 2024 menggunakan kombinasi TF-IDF dengan Random Forest untuk mendeteksi promosi judi online di Twitter berbahasa Indonesia dan mencapai akurasi 97,87%. Namun, pendekatan ini masih memiliki keterbatasan dalam memahami konteks semantik yang kompleks [2]. Selain itu, terdapat penelitian yang membandingkan algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam analisis sentimen terkait judi online, dengan SVM mencapai akurasi tertinggi sebesar 98% [3]. Meskipun demikian, model-model tradisional ini belum mampu menangkap nuansa bahasa dan variasi linguistik yang terus berkembang.

Perkembangan teknologi *Natural Language Processing* (NLP), khususnya model berbasis Transformer, telah

membuka peluang baru dalam deteksi konten berbahaya secara otomatis. Model BERT (*Bidirectional Encoder Representations from Transformers*) telah membawa perubahan signifikan dalam bidang NLP dengan kemampuannya memahami konteks kata. Pada penelitian 2023 mengembangkan model berbasis BERT untuk mendeteksi tanda-tanda perjudian bermasalah dalam forum daring berbahasa Jerman dengan akurasi 81% dan presisi 95% [4]. Namun, BERT memiliki kelemahan dalam hal kebutuhan sumber daya komputasi yang besar dan waktu inferensi yang relatif lama.

Mengatasi keterbatasan tersebut, dikembangkan DistilBERT, sebuah varian BERT yang lebih ringan dan efisien. Dilihat pada penelitian tahun 2022 membandingkan performa BERT dan DistilBERT dalam analisis sentimen Twitter terkait Covid-19 dan menemukan bahwa DistilBERT unggul dengan akurasi 97% serta kecepatan pemrosesan yang lebih tinggi, dibandingkan BERT yang hanya mencapai akurasi 87% [5]. DistilBERT mampu mempertahankan sekitar 97% performa BERT dengan hanya 40% dari jumlah parameter [6]Di sisi lain, DeBERTa (*Decoding-enhanced BERT with Disentangled Attention*) menawarkan pendekatan yang berbeda dengan mekanisme disentangled attention yang lebih canggih. Tahun 2024 dilakukan penelitian yang menunjukkan bahwa DeBERTa secara signifikan lebih unggul dibandingkan BERT dalam klasifikasi teks politik, terutama setelah dilakukan *fine-tuning* yang tepat[7].

Pendekatan *transfer learning* dan *fine-tuning* akan digunakan untuk mengoptimalkan performa kedua model, memungkinkan model untuk memanfaatkan pengetahuan bahasa yang telah dipelajari dari corpus besar, kemudian disesuaikan dengan karakteristik unik dari konten promosi judi online [8]. Selain itu, penelitian ini juga mengintegrasikan analisis *word cloud* untuk mengidentifikasi pola kata kunci yang dominan dalam promosi judi online, sehingga dapat memberikan wawasan lebih dalam tentang karakteristik linguistik yang digunakan oleh para pelaku [9].

Hasil penelitian ini diharapkan dapat memberikan kontribusi teoritis dalam memperkaya literatur ilmiah di bidang NLP, khususnya dalam aplikasi model Transformer untuk deteksi konten berbahaya. Secara praktis, penelitian ini dapat diimplementasikan dalam pengembangan sistem deteksi otomatis untuk moderasi konten di platform media sosial, dan membantu melindungi pengguna. Dengan meningkatkan efektivitas dan efisiensi dalam mengidentifikasi dan memfilter konten berbahaya secara real-time, sistem ini dapat mengurangi ketergantungan pada moderasi manual yang memerlukan banyak sumber daya.

II. METODE

A. Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari platform Kaggle dengan judul "Deteksi Judi Online" yang dapat diakses melalui tautan https://www.kaggle.com/datasets/yaemico/deteksi-judi-online dengan lisensi *Apache 2.0*. Data dikumpulkan dari kolom komentar pada video YouTube berjudul "LIVE Wayang Jogja Night Carnival #9" dalam rangka Perayaan HUT ke-268 Kota Yogyakarta yang diunggah pada kanal resmi penyelenggara acara. Periode pengambilan data dilakukan pada tahun 2024, bertepatan dengan waktu penyelenggaraan acara dan aktivitas interaksi penonton melalui kolom komentar siaran langsung tersebut.

Dataset tersebut telah melalui proses pelabelan oleh pembuat dataset yang terdiri dari 6.350 komentar dengan distribusi yang seimbang. Kelas "Bukan Promosi" (label 0) sebanyak 3.163 komentar dan kelas "Promosi" (label 1) sebanyak 3.187 komentar. Distribusi *balanced* dataset ini penting untuk menghindari bias dalam pelatihan model dan memastikan evaluasi yang adil. Karakteristik dataset menggunakan bahasa Indonesia dengan variasi dialek, slang, dan singkatan media sosial, serta kompleksitas linguistik dari formal hingga informal dengan penggunaan emoji, simbol khusus, dan teknik penyamaran kata. Namun demikian, penelitian ini memiliki beberapa keterbatasan. Cakupan data masih terbatas pada satu kanal YouTube sehingga belum merepresentasikan variasi bahasa pengguna media sosial secara luas. Selain itu, penggunaan bahasa informal dan campuran (code-mixed) dapat menurunkan performa model karena belum dioptimalkan untuk teks jenis tersebut.

Dengan deskripsi komprehensif ini, penelitian memiliki fondasi kuat dalam kualitas, reliabilitas, dan validitas data, sehingga hasil analisis dapat dipertanggungjawabkan secara ilmiah dan memberikan kontribusi signifikan dalam pengembangan sistem deteksi konten berbahaya di media sosial.

Dari sisi etika penelitian, seluruh data yang digunakan berasal dari komentar publik di YouTube tanpa mengakses informasi pribadi pengguna. Dengan demikian, penelitian ini mematuhi prinsip etika penggunaan data publik serta memastikan bahwa proses pengumpulan dan pengolahan data tidak melanggar privasi maupun hak digital pengguna.

B. Pre-processing Data

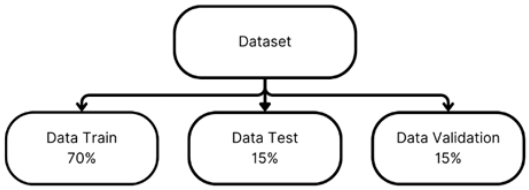
Tahapan *pre-processing* data merupakan langkah penting untuk mempersiapkan teks mentah agar sesuai dengan kebutuhan model dan meningkatkan kualitas pembelajaran. Proses ini bertujuan untuk menghilangkan *noise*, menstandarkan format, serta memastikan data yang digunakan bersih dan konsisten. Tahapan ini dilakukan secara sistematis, yaitu:

a. *case folding*, yaitu mengubah seluruh teks menjadi huruf kecil (*lowercase*) untuk menjaga keseragaman dan mengurangi variasi kata akibat perbedaan penulisan.
b. Pembersihan URL yang sering digunakan dalam komentar untuk mempromosikan situs judi online, baik dalam bentuk lengkap maupun tautan yang disamarkan.
c. Penghapusan karakter khusus dan tanda baca yang tidak memiliki arti penting, karena model *transformer* seperti BERT sudah dapat memahami struktur kalimat tanpa bergantung pada tanda baca eksplisit.

d. Normalisasi kata tidak baku, mengingat bahasa Indonesia di media sosial sering menggunakan kata singkatan, *slang*, atau bentuk tidak baku. Normalisasi dilakukan dengan membangun kamus (*dictionary*) pemetaan dari kata tidak baku ke bentuk bakunya agar data menjadi lebih konsisten.
e. Terakhir, dilakukan *filtering* terhadap teks kosong, komentar terlalu pendek (kurang dari tiga karakter atau dua kata), serta penghapusan komentar duplikat untuk menghindari bias pada proses pelatihan.

Dengan tahapan ini, dataset yang dihasilkan menjadi lebih bersih, terstruktur, dan siap digunakan untuk proses *tokenization* serta pelatihan model berbasis *transformer* seperti DistilBERT dan DeBERTa [10].

Setelah melalui tahap pra-pemrosesan tersebut, dataset kemudian dibagi ke dalam tiga subset, yaitu data latih (*training set*), data validasi (*validation set*), dan data uji (*test set*). Proses pembagian dilakukan secara acak (*random shuffle*) untuk menjaga agar setiap subset tetap merepresentasikan karakteristik keseluruhan dataset. Dengan demikian, model dapat dilatih secara optimal serta dievaluasi secara objektif menggunakan data yang belum pernah dipelajari sebelumnya [9]. Distribusi pembagian dataset ditunjukan pada Gambar 1.



Gambar 1. Pembagian Data [11]

C. *Representasi Teks*

Dalam penelitian ini, proses representasi teks sepenuhnya dilakukan dengan memanfaatkan *word embedding* berbasis model Transformer dan tidak dilakukan secara manual. *Word embedding* adalah representasi kata dalam bentuk vektor numerik yang digunakan dalam pemrosesan bahasa alami untuk menangkap makna semantik dan hubungan antar kata yang dapat dimengerti oleh komputer. Proses representasi teks diawali dengan tokenisasi yang bertujuan menghasilkan token dari teks mentah menggunakan metode *WordPiece* untuk DistilBERT dan *Byte-Pair Encoding* (BPE) untuk DeBERTa. Metode WordPiece memecah kata menjadi unit-unit lebih kecil (*subword*) jika kata tersebut tidak ditemukan dalam kosakata tokenizer, dengan menambahkan tanda ## pada awal token untuk menunjukkan kelanjutan dari kata sebelumnya, sehingga memungkinkan model menangani kata-kata yang jarang atau tidak dikenal (*out-of-vocabulary words*).

Sementara itu, metode BPE pada DeBERTa menggunakan pendekatan *byte-level* yang memperlakukan setiap karakter termasuk spasi sebagai bagian dari token, dengan tanda khusus "Ġ" yang muncul di awal token untuk menandai adanya spasi sebelum kata tersebut, sehingga lebih tangguh dalam menangani variasi teks informal, kesalahan ketik, serta karakter non-standar yang sering muncul pada komentar media sosial. Tokenizer secara otomatis menambahkan token khusus seperti [CLS] untuk menandai awal teks dan [SEP] untuk memisahkan atau mengakhiri teks dalam tugas klasifikasi.

Setelah tokenisasi, setiap token dikonversi menjadi ID angka berdasarkan kamus tokenizer yang telah didefinisikan sebelumnya. ID token yang telah berbentuk tensor selanjutnya dimasukkan ke dalam *embedding layer* model, yang merupakan matriks besar yang memetakan setiap token numerik menjadi vektor kontinu berdimensi 768. Batas maksimum panjang urutan (*max sequence length*) sebesar 128 token. Nilai-nilai pada vektor *embedding* diperoleh dari hasil *pre-training* model pada korpus teks berskala besar. Vektor *embedding* kemudian diproses lebih lanjut melalui lapisan-lapisan transformer dalam model DistilBERT dan DeBERTa yang akan menangkap konteks dan hubungan antar token secara mendalam menggunakan *mechanism attention*[12].

D. *Distillation BERT*

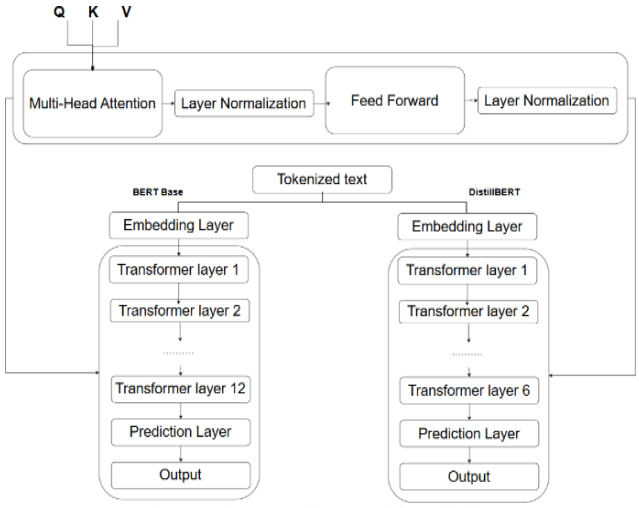
DistilBERT adalah varian ringan dari BERT yang dikembangkan untuk memperkecil ukuran model sekaligus mempertahankan sebagian besar kemampuannya. Pelatihannya menggunakan metode *knowledge distillation*, yaitu proses ketika model kecil meniru perilaku model besar. Dengan metode ini, DistilBERT memiliki parameter 40% lebih sedikit dan kecepatan inferensi 60% lebih cepat, tetapi tetap mempertahankan 97% performa BERT untuk berbagai tugas NLP. Arsitektur DistilBERT menggunakan enam lapisan transformer dengan mekanisme *self-attention* yang tetap mempertahankan kemampuan memahami hubungan kontekstual antar kata[6]

E. *Decoding*-enhanced *BERT with Disentangled Attention*

DeBERTa digunakan untuk meningkatkan kinerja BERT melalui tiga inovasi utama:

a. *Disentangled attention mechanism*, yang memisahkan representasi konten (*content embedding*) dan posisi (*position embedding*) dalam perhitungan atensi
b. *Enhanced mask decoder*, yang memperbaiki pemahaman konteks saat pemodelan *masked language*
c. *Absolute + relative position encoding*, yang memadukan informasi posisi absolut dan relatif untuk memperkuat pemahaman urutan kata. Perbaikan ini memungkinkan DeBERTa mencapai hasil unggul di berbagai benchmark NLP dengan efisiensi yang tetap kompetitif [13]

F. *Pemodelan DistilBERT*



Gambar 1. Diagram Alir Pemodelan DistilBERT

Pada penelitian ini diawali dengan proses *fine-tuning* terhadap data komentar YouTube yang telah melalui tahap *pre-processing* dan representasi teks berbasis *word embedding*. Setiap lapisan transformer memanfaatkan *multi-head self-attention* untuk menangkap hubungan kata dari berbagai konteks, diikuti oleh layer *normalization* untuk menjaga stabilitas distribusi nilai, dan *feed-forward network* untuk memperkuat representasi sebelum diteruskan ke lapisan berikutnya.

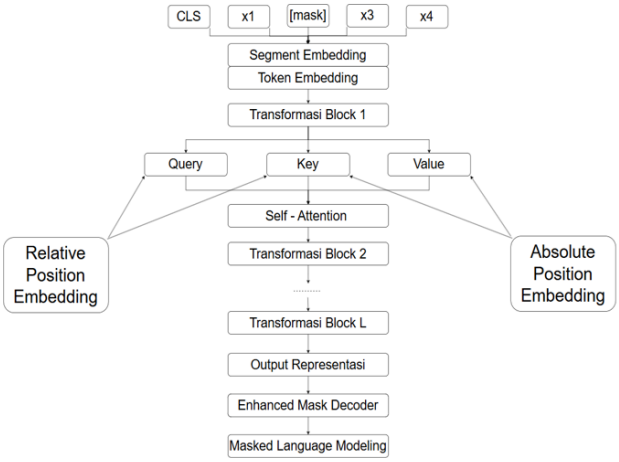
Selama proses pelatihan, DistilBERT menggunakan tiga komponen *loss function* utama.

- Pertama, *masked language modeling* (MLM) yang melatih model untuk memprediksi kata yang ditutupi dalam sebuah kalimat.
- Kedua, *distillation loss* yang memanfaatkan keluaran BERT sebagai acuan pembelajaran sehingga model siswa dapat meniru perilaku model guru.
- Ketiga, *cosine distance loss* yang menjaga keselarasan arah vektor status tersembunyi antara kedua model.

Penggabungan ketiga komponen ini memungkinkan DistilBERT mempertahankan performa tinggi meskipun memiliki arsitektur yang lebih ringan. Dalam penelitian ini, parameter pelatihan seperti *learning rate, jumlah epoch,* dan *batch size* disesuaikan untuk menjaga stabilitas pembelajaran serta menghindari *overfitting* atau *underfitting* [6]

G. Pemodelan DeBERTa

Model ini dipilih karena keunggulannya mempertahankan informasi hubungan antar kata secara lebih konsisten, bahkan pada kalimat yang memiliki struktur kompleks atau mengandung makna tersirat. Dengan arsitektur yang dirancang untuk memaksimalkan penangkapan pola bahasa, DeBERTa sangat relevan dalam mendeteksi promosi judi online yang sering menggunakan kata-kata pengganti, sinonim, atau frasa yang dimodifikasi agar lolos dari deteksi manual. Dalam proses pelatihannya, model DeBERTa sama seperti model DistilBERT. Bedanya pada representasi, model ini diproses melalui blok transformer yang menggabungkan informasi konteks dari berbagai posisi dalam kalimat. Mekanisme pengolahan ini memungkinkan model memahami hubungan sintaksis dan semantik secara simultan, sehingga representasi akhir mengandung informasi makna kata sekaligus posisinya dalam teks [13]



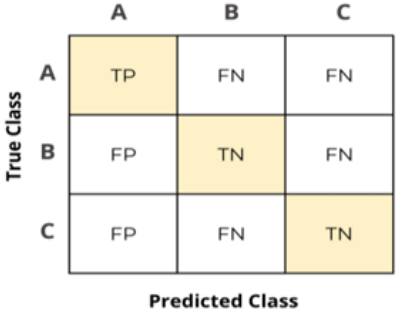
Gambar 2. Diagram Alir Pemodelan DeBERTa

Untuk memastikan hasil perbandingan yang adil dengan DistilBERT, parameter pelatihan seperti *learning rate*, jumlah *epoch*, dan *batch size* dibuat sama. Dengan pendekatan ini, diharapkan DeBERTa mampu memberikan kinerja optimal pada deteksi teks promosi judi online, baik dari sisi akurasi maupun kemampuan mengidentifikasi pola bahasa yang tersembunyi.

H. Evaluasi Model

Proses evaluasi ini dilakukan pada data validasi untuk memberikan gambaran tentang performa model sebelum diujikan pada data yang benar-benar baru. Dengan evaluasi ini, performa model dapat direfleksikan lebih akurat tanpa bias dari data yang digunakan untuk pelatihan atau penyesuaian model. Metrik evaluasi yang digunakan mencakup:

1. *Confusion Matrix:* Matriks ini memberikan total prediksi yang benar dan salah pada masing-masing kelas.



Gambar 3. Konsep Confusion Matrix[14]

Dimana label-label pada confusion matrix meliputi:

- *True Positive* (TP), yaitu prediksi positif yang sesuai
- *True Negative* (TN), prediksi negatif yang benar
- *False Positive* (FP), prediksi positif yang salah
- *False Negative* (FN), prediksi negatif yang keliru.

Berdasarkan confusion matrix tersebut, performa model kemudian dievaluasi menggunakan sejumlah metrik klasifikasi umum yaitu *precision*, *recall*, *F1-score*, dan *accuracy* untuk setiap kelas, di mana *precision* menunjukkan ketepatan prediksi positif yang benar-benar positif, *recall* mengukur kemampuan model untuk menangkap seluruh data positif yang sebenarnya, dan *F1-score* merepresentasikan keseimbangan antara *precision* dan *recall*, yang sangat berguna ketika data memiliki distribusi kelas yang tidak seimbang [15].

2. *ROC AUC Score:* Area under the ROC curve (ROC AUC) adalah metrik yang mengukur kemampuan model dalam memisahkan kelas positif dan negatif pada berbagai ambang batas. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki performa yang baik dalam membedakan antara kelas promosi dan bukan promosi. ROC AUC digunakan karena mampu menggambarkan performa model secara lebih umum tanpa bergantung pada ambang batas tertentu [11].

I. Environment Setup

Eksperimen dilakukan menggunakan platform Google Colab dengan spesifikasi perangkat keras GPU NVIDIA Tesla T4 yang memiliki memori 15,36 GB, dan CUDA version 12.4. Sistem memiliki RAM sebesar 13,29 GB dengan lingkungan Python 3.12.12. Proses implementasi dan pelatihan model dilakukan menggunakan framework *PyTorch* sebagai backend utama dari library Transformers. Selain itu, digunakan pula pustaka pendukung meliputi *PyTorch* 2.1.0, Transformers 4.31.0, Pandas, NumPy, *Scikit-learn*, *Imbalanced-learn* (imblearn), NLTK, *Matplotlib*, *Seaborn*, dan *WordCloud*, yang berperan dalam proses pre-processing, visualisasi, dan evaluasi model.

III. HASIL DAN PEMBAHASAN

A. Pre-processing

Setelah dataset melalui tahap pembersihan, akan dilakukan penyeimbangan menggunakan teknik undersampling. sehingga diperoleh distribusi yang seimbang pada kedua label, masing-masing sebanyak 3.163 data. Selanjutnya, dataset tersebut dibagi menjadi tiga bagian:

- data latih (70%) sebanyak 4.428 data
- data validasi (15%) sebanyak 948 data
- data uji (15%) sebanyak 950 data

yang digunakan untuk memastikan performa model dapat dievaluasi secara adil pada data yang tidak terlihat selama proses pelatihan.

Tokenisasi berfungsi menyiapkan teks ke dalam format yang dapat diproses oleh model. Pada tahap ini, teks dipecah menjadi token numerik dengan tetap menjaga struktur kata maupun sub-kata. Langkah tersebut menjamin data tetap rapi, seragam, serta siap digunakan dalam pelatihan model sehingga performa dapat meningkat.

B. Pemodelan DistilBERT dan DeBERTa

Pemodelan dimulai dengan menetapkan beberapa *hyperparameter* yang dipakai, yaitu:

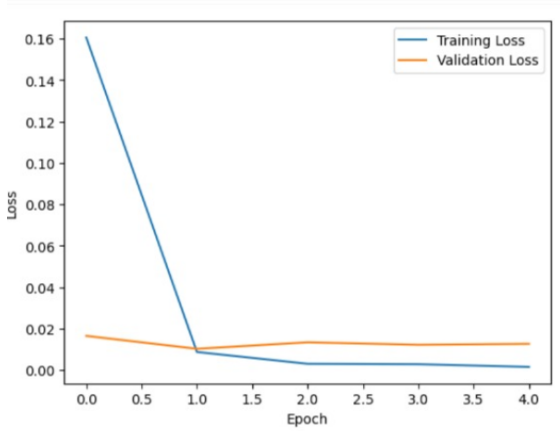
TABEL I.
HYPERPARAMETER MODEL

Hyperparamater	Nilai
Learning rate	1e-5
Batch size	32
Epoch	5
Weight decay	0.2

Proses fine-tuning merupakan tahapan krusial dalam mengadaptasi model pre-trained untuk tugas klasifikasi spesifik, yaitu deteksi promosi judi online. Pada penelitian ini, fine-tuning dilakukan pada kedua model DistilBERT dan DeBERTa dengan menggunakan konfigurasi hyperparameter yang telah dioptimalkan untuk mencapai keseimbangan antara kecepatan konvergensi dan kemampuan generalisasi model. Konfigurasi hyperparameter yang digunakan meliputi:

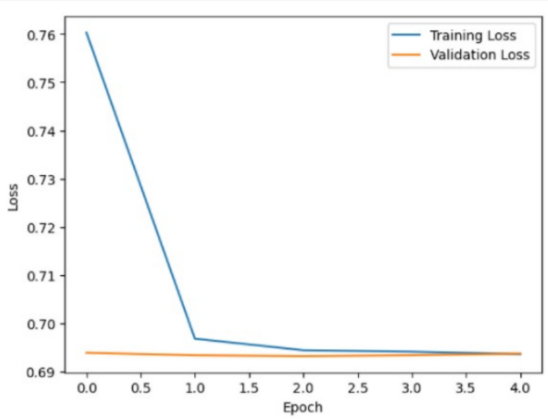
a. *learning rate* sebesar 2×10^{-5} yang mengontrol seberapa besar langkah pembaruan bobot dalam setiap iterasi pelatihan, sehingga mempengaruhi kecepatan dan stabilitas konvergensi model.
b. *Batch size* ditentukan sebesar 32, yang berarti pada setiap iterasi model memproses 32 sampel data sebelum melakukan pembaruan bobot.
c. Jumlah *epoch* ditetapkan sebanyak 5, yang menunjukkan berapa kali model akan melalui seluruh dataset pelatihan secara lengkap tanpa memerlukan pelatihan yang terlalu panjang.
d. *Weight decay* ditetapkan sebesar 0,01 sebagai teknik regularisasi untuk mencegah overfitting dengan menambahkan penalti pada besarnya nilai bobot model

Data *train* digunakan agar model dapat mempelajari perbedaan antara kelas promosi dan bukan promosi melalui penyesuaian bobot berdasarkan *Training Loss*. Sementara itu, data *validasi* dievaluasi pada akhir setiap *epoch* untuk memantau kinerja model, mendeteksi potensi *overfitting*, serta memastikan model tidak sekadar mengingat data latih.



Gambar 5. Training Data DistilBERT

Gambar 5 menunjukkan kurva *training loss* dan *validation loss* dari proses pelatihan model DistilBERT. Nilai *training loss* awalnya tinggi ~0,16 lalu menurun tajam hingga mendekati nol sejak epoch pertama, sedangkan validation loss tetap stabil di kisaran rendah 0,01–0,02. Pola ini menunjukkan kemampuan belajar yang cepat, namun terdapat indikasi awal overfitting karena perbedaan cukup jelas antara training loss dan validation loss.



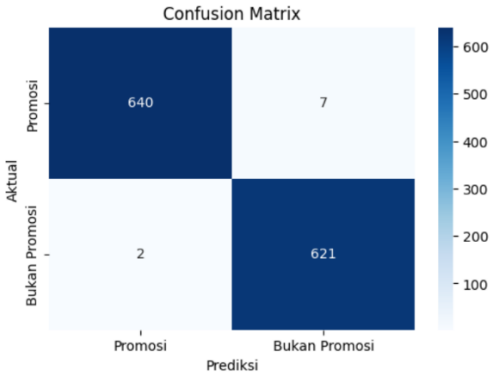
Gambar 6. Training Data DeBERTa

Gambar 6 menunjukkan kurva *training loss* dan *validation loss* dari proses pelatihan model DeBERTa. Nilai *training loss* awalnya lebih tinggi ~0,76 kemudian menurun secara bertahap hingga sekitar 0,69 pada epoch kedua, dengan *validation loss* yang stabil di kisaran ~0,69 tanpa fluktuasi berarti. Hal ini menandakan bahwa DeBERTa mampu belajar secara lebih konsisten tanpa gejala overfitting, sehingga memiliki kemampuan generalisasi yang lebih baik dibandingkan DistilBERT.

Berdasarkan hasil kurva *loss*, menunjukkan bahwa DistilBERT mengalami penurunan *training loss* yang sangat cepat hingga mendekati nol pada epoch kedua, sedangkan *validation loss* stabil di kisaran ~0,01, pola ini menandakan efisiensi pembelajaran namun dengan potensi overfitting karena selisih cukup besar antara *training* dan *validation loss*. Sebaliknya, DeBERTa menampilkan penurunan *training loss* yang lebih bertahap dari ~0,76 menjadi ~0,69 dengan *validation loss* yang konsisten mengikuti pola serupa, sehingga meskipun membutuhkan waktu lebih lama untuk konvergensi, model ini menunjukkan stabilitas yang lebih baik serta kemampuan generalisasi yang relatif lebih kuat dibandingkan DistilBERT.

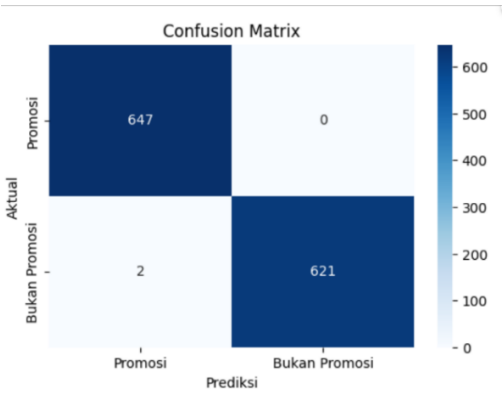
C. Performa Model

1. *Confusion Matrix:*



Gambar 7. Confussion Matrix Model DistilBERT

Berdasarkan *confusion matrix* di Gambar 7, model DistilBERT memperlihatkan performa yang juga sangat baik, dengan 640 data "Promosi" terklasifikasi benar dan 7 data salah diklasifikasikan sebagai "Bukan Promosi". Untuk kelas "Bukan Promosi", terdapat 621 data yang diprediksi benar dan hanya 2 data yang keliru diprediksi sebagai "Promosi". Walaupun akurasi dan presisi masih tinggi, terlihat bahwa DistilBERT sedikit lebih banyak melakukan kesalahan klasifikasi dibandingkan DeBERTa, khususnya dalam membedakan komentar promosi.



Gambar 8. Confussion Matrix Model DeBERTa

Berdasarkan Gambar 8, hasil confusion matrix model DeBERTa menunjukkan performa yang sangat baik dengan 647 data "Promosi" terklasifikasi benar dan hanya 2 data "Bukan Promosi" yang salah diklasifikasikan sebagai "Promosi". Tidak ada kesalahan klasifikasi pada kelas "Promosi" ke kelas “bukan promosi”, serta 621 data "Bukan Promosi" berhasil diprediksi dengan tepat. Hasil ini

mencerminkan tingkat akurasi yang sangat tinggi, presisi mendekati sempurna, dan kemampuan generalisasi yang kuat dalam membedakan komentar promosi judi online dan bukan promosi.

2. *Error Analysis:* Setelah dilakukan evaluasi menggunakan *confusion matrix*, tahap berikutnya adalah melakukan error analysis untuk mengidentifikasi jenis dan penyebab kesalahan klasifikasi yang terjadi pada model. Analisis ini bertujuan untuk memberikan pemahaman yang lebih mendalam. Berdasarkan hasil evaluasi, sebagian kecil data masih diklasifikasikan secara keliru, baik dalam bentuk *False Positive* (FP) maupun *False Negative* (FN). Kesalahan tersebut dapat disebabkan oleh kesamaan struktur, ambiguitas konteks, serta pengaruh tahap pre-processing yang mengubah bentuk asli kata sehingga maknanya sulit dikenali oleh model.

Oleh karena itu, bagian ini membahas secara rinci contoh-contoh kesalahan yang terjadi pada model DistilBERT dan DeBERTa untuk memahami karakteristik pola teks yang paling memengaruhi performa deteksi promosi judi online.

TABEL II.
KESALAHAN PADA MODEL DISTILBERT

Message	Label Asli	Label Prediksi	Analisis
d3p0 100 jd 2jt buruan gas	1	0	Model gagal mengenali singkatan numerik seperti d3p0 (deposit)
freebet 100rb wisdomtoto	1	0	nama situs tersamar "wisdomtoto" karena bentuknya tidak umum di korpus pelatihan, sehingga model gagal mengidentifikasi konteks promosi
ketik di google wisdomtoto	1	0	Frasa perintah sederhana tanpa kata eksplisit seperti *slot*, *judi* membuat model salah mendeteksinya
100rb gratis wisdomtoto	1	0	Singkatan "rb" menyebabkan tokenisasi ambigu, model tidak mampu menangkap konsep bonus promosi
freebet 100rb wisdomtoto	1	0	Duplikasi pola singkat seperti ini sering dianggap *noise* oleh model karena tidak ada variasi sintaksis.
ketik di googlewisdomtoto	1	0	Penulisan kata tanpa spasi setelah pembersihan menyebabkan token "googlewisdomtoto" tidak dikenal
100rb gratis bonus tanpa deposit	1	0	Tidak mencantumkan nama situs, sehingga model gagal mengidentifikasi konteks promosi meskipun secara semantik mengandung ajakan deposit.
jangan coba main slot temen	0	1	Kata "main" dan "slot" diasosiasikan kuat dengan konteks judi online.
link gacor	0	1	Kata "gacor" sering muncul pada data promosi, tapi konteksnya berbeda.

TABEL III.
KESALAHAN PADA MODEL DEBERTA

Message	Label Asli	Label Prediksi	Analisis
jangan coba main slot temen	0	1	Kata "main" dan "slot" diasosiasikan kuat dengan konteks judi online.
link gacor	0	1	Kata "gacor" sering muncul pada data promosi, tapi konteksnya berbeda.

3. *ROC AUC Score:* Kedua model mencapai ROC AUC score sebesar 0.99 pada data uji, menunjukkan kemampuan model untuk membedakan antara kelas promosi dan bukan promosi dengan baik, yang penting untuk aplikasi deteksi promosi judi online.

4. *Analisis Overfitting:*

TABEL IV.
CROSS VALIDATION DISTILBERT

Fold	Akurasi	Presisi	Recall	F1-Score
1	99.7%	99.7%	99.7%	99.7%
2	100%	100%	100%	100%
3	100%	100%	100%	100%
4	99.7%	99.7%	99.7%	99.7%
5	99.9%	99.9%	99.9%	99.9%
Rata-Rata	99.8%	99.8%	99.8%	99.8%

Untuk menghindari kecurigaan overfitting mengingat akurasi yang sangat tinggi (≈99%), dilakukan *k-fold cross-validation* dengan $k = 5$ pada model DistilBERT. Hasil pada Tabel IV menunjukkan performa yang sangat konsisten di seluruh fold dengan akurasi rata-rata 99,8% dan standar deviasi hanya 0,14%. Nilai precision, recall, dan F1-score juga menunjukkan variabilitas rendah dengan standar deviasi <0,2%. Konsistensi ini mengonfirmasi bahwa performa tinggi bukan merupakan hasil overfitting, melainkan mencerminkan kemampuan model dalam mengenali pola promosi judi online. Distribusi data yang seimbang dan karakteristik tekstual yang distinktif antara kelas "Promosi" dan "Bukan Promosi" memungkinkan model untuk belajar secara efektif tanpa mengalami overfitting yang signifikan.

TABEL V.
HASIL CROSS VALIDATION DEBERTA

Fold	Akurasi	Presisi	Recall	F1-Score
1	99.9%	99.9%	99.9%	99.9%
2	99.9%	99.9%	99.9%	99.9%
3	99.9%	99.9%	99.9%	99.9%
4	99.9%	99.9%	99.9%	99.9%
5	99.9%	99.9%	99.9%	99.9%
Rata-Rata	99.9%	99.9%	99.9%	99.9%

Tabel V menunjukkan hasil pengujian model menggunakan metode 5-*Fold Cross Validation* menunjukkan performa yang sangat konsisten dan tinggi pada setiap fold. Nilai akurasi, presisi, recall, dan F1-score pada masing-masing fold berada di kisaran 99.9%, yang berarti model mampu melakukan klasifikasi dengan tingkat kesalahan yang sangat rendah. Konsistensi nilai antar fold menunjukkan bahwa model memiliki stabilitas dan kemampuan generalisasi yang baik terhadap data uji. Secara keseluruhan, rata-rata hasil evaluasi sebesar 99.9% pada semua metrik menunjukkan bahwa model mampu mendeteksi kategori data dengan tingkat ketepatan yang hampir sempurna.

5. Perbandingan Performa Model:

TABEL VI.
PERBANDINGAN PERFORMA MODEL

Model	Akurasi	Presisi	Recall	F1-score
SVM+TF-IDF	99.89%	99.79%	100%	99.9%
DeBERTa	99.84%	99.7%	100%	99.8%
DistilBERT	99.29%	99.7%	98.9%	99.3%

Berdasarkan hasil evaluasi, ketiga model yang diuji menunjukkan performa yang sangat baik dengan akurasi di atas 99%. Model DeBERTa mencapai akurasi 99,84%, precision 99,7%, recall 100%, dan F1-score 99,8%, menandakan bahwa model ini mampu mengklasifikasi komentar promosi secara sempurna tanpa kehilangan data positif. Sementara itu, DistilBERT memperoleh akurasi 99,29%, precision 99,7%, recall 98,9%, dan F1-score 99,3%. Meskipun masih sangat tinggi, performa DistilBERT sedikit lebih rendah pada aspek recall, yang menunjukkan adanya beberapa komentar promosi yang tidak terdeteksi. Sebagai baseline pembanding, digunakan model SVM+TF-IDF dengan hasil akurasi 99,89%, precision 99,79%, recall 100%, dan F1-score 99,9%. Hasil ini menunjukkan bahwa model klasik masih mampu bersaing secara kuantitatif dengan model transformer.

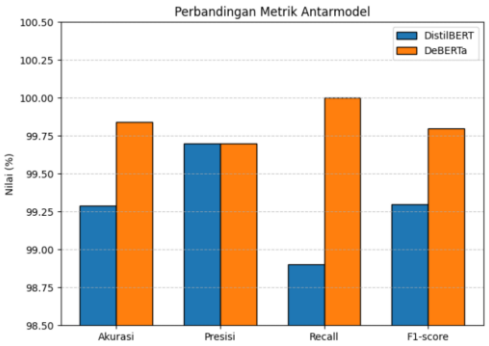
Secara arsitektural, model DeBERTa menunjukkan performa yang lebih unggul dibandingkan DistilBERT dalam tugas deteksi promosi judi online. Keunggulan ini berasal dari mekanisme *disentangled attention* dan *enhanced mask decoder* pada arsitektur DeBERTa, yang memungkinkan model memisahkan representasi posisi dan konten kata sehingga mampu memahami konteks kalimat dengan lebih akurat, terutama pada teks berstruktur kompleks seperti komentar media sosial.

Dari sisi efisiensi komputasi, DistilBERT memiliki sekitar 66 juta parameter, sedangkan DeBERTa base mencapai 140 juta parameter. Perbedaan ini berdampak pada waktu pelatihan dan kecepatan inferensi. DistilBERT mampu menyelesaikan proses pelatihan dalam waktu sekitar 5 menit dengan penggunaan memori GPU yang lebih rendah, menjadikannya lebih efisien untuk implementasi pada perangkat dengan sumber daya terbatas. Sebaliknya, DeBERTa memerlukan waktu sekitar 11 menit karena arsitekturnya yang lebih kompleks, namun menghasilkan performa yang lebih tinggi.

Dengan demikian, DistilBERT lebih cocok digunakan untuk kebutuhan *real-time* atau aplikasi ringan yang menuntut efisiensi waktu dan sumber daya, sedangkan DeBERTa lebih optimal untuk skenario yang menekankan akurasi tinggi dan pemahaman konteks mendalam. Sementara itu, keberadaan model SVM+TF-IDF sebagai baseline memperkuat analisis perbandingan ini, menunjukkan bahwa meskipun transformer unggul dalam pemahaman semantik, model klasik masih relevan sebagai pendekatan yang efisien dan kompetitif secara performa.

D. Visualisasi Model

1. Perbandingan Metrik Antarmodel:



GAMBAR 9. PERBANDINGAN METRIK ANTARMODEL

Gambar tersebut memperlihatkan perbandingan nilai metrik kinerja antara model DistilBERT dan DeBERTa berdasarkan hasil pengujian. Secara umum, model DeBERTa menunjukkan performa yang lebih unggul dibandingkan DistilBERT pada seluruh metrik evaluasi, terlihat pada nilai recall yang paling signifikan, di mana DeBERTa mampu mendeteksi komentar promosi dengan tingkat keberhasilan lebih tinggi dibandingkan DistilBERT.

Hal ini menunjukkan bahwa DeBERTa memiliki kemampuan yang lebih baik dalam memahami konteks kalimat yang kompleks dan variasi bahasa pada komentar YouTube. Sementara itu, meskipun DistilBERT menunjukkan performa yang cukup baik, model ini cenderung menghasilkan lebih banyak kesalahan pada data dengan struktur kalimat yang ambigu. Dengan demikian,

visualisasi ini mengonfirmasi bahwa arsitektur DeBERTa lebih efektif untuk tugas klasifikasi teks yang membutuhkan pemahaman konteks mendalam seperti deteksi promosi judi online

2. *Analisis Word Cloud:*

Gambar 10. Hasil Analisis Word Cloud

Word cloud pada teks promosi judi online memperlihatkan dominasi kata-kata yang berhubungan dengan penawaran bonus seperti *freebet*, *gratis*, serta instruksi akses website (*ketik google*, *rb*). Sebaliknya, word cloud pada teks bukan promosi lebih banyak berisi kata-kata umum dan percakapan sehari-hari, seperti *selamat ulang*, *jogja*, dan *tugu*. Perbedaan ini menunjukkan bahwa fitur teks promosi judi online memiliki pola kosakata yang lebih spesifik dan berorientasi pada ajakan promosi, sementara teks bukan promosi lebih bervariasi dan kontekstual.

E. *Analisis Linguistik*

Salah satu tantangan utama dalam deteksi promosi judi online adalah kemampuan pelaku dalam menggunakan ambiguitas bahasa dan istilah terselubung untuk menghindari sistem deteksi. Penelitian ini menganalisis secara mendalam bagaimana model menangani berbagai strategi penyamaran linguistik yang umum digunakan dalam promosi judi online, termasuk substitusi karakter, singkatan kreatif, dan penggunaan frasa ambigu.

1. *Pola Linguistik Promosi Judi Online:* Berdasarkan analisis korpus data, promosi judi online menunjukkan karakteristik linguistik yang khas dan berbeda signifikan dari komentar normal.

Pada analisis word cloud memperlihatkan dominasi kata-kata yang berhubungan dengan penawaran insentif seperti freebet, gratis, bonus, dan deposit, Sebaliknya, komentar non-promosi didominasi oleh kata-kata percakapan sehari-hari seperti selamat ulang tahun, jogja, dan tugu, yang menunjukkan konteks interaksi sosial normal pada acara komunitas. Melalui error analysis yang dilakukan pada model, teridentifikasi beberapa kategori strategi penyamaran linguistik yang digunakan pelaku promosi judi online:

a. Penggantian huruf dengan angka atau simbol untuk menghindari deteksi kata kunci, seperti "d3p0" untuk "deposit", "b0nus" untuk "bonus", dan "0nline" untuk "online". Teknik ini memanfaatkan kesamaan visual antara karakter alfanumerik.

b. Penggunaan singkatan kreatif yang tidak umum dalam kamus bahasa formal, seperti "rb" untuk "ribu", "jt" untuk "juta", dan "gas" sebagai ajakan untuk segera bertindak. Singkatan ini sering dikombinasikan dengan angka untuk membentuk frasa promosi seperti "100rb gratis".

c. Strategi menyatukan beberapa kata menjadi satu kesatuan tanpa pemisah, seperti "googlewisdomtoto" atau "ketikdigoogle", yang bertujuan untuk mengaburkan struktur sintaksis dan menyulitkan sistem deteksi berbasis keyword.

d. Penggunaan kalimat yang tidak secara eksplisit menyebutkan judi atau situs tertentu, namun mengandung ajakan tersirat seperti "ketik di google wisdomtoto" atau "freebet 100rb", yang bergantung pada pemahaman konteks untuk mengidentifikasi maksud promosi.

2. *Mekanisme Tokenisasi dalam Menangani Ambiguitas:* Perbedaan fundamental antara DistilBERT dan DeBERTa terletak pada metode tokenisasi yang digunakan, yang secara signifikan memengaruhi kemampuan model dalam menangani teks informal dan ambigu.

a. Model DistilBERT menggunakan metode WordPiece yang memecah kata menjadi subword units berdasarkan kosakata yang telah dipelajari selama pre-training. Metode ini menambahkan prefix "##" untuk menandai token yang merupakan kelanjutan dari kata sebelumnya, di mana model mencoba mengenali komponen kata yang familiar. Namun, ketika menghadapi substitusi numerik seperti "d3p0", tokenizer cenderung memecahnya menjadi token individual, yang kehilangan makna semantik dari kata asli "deposit". Hal ini menjelaskan mengapa DistilBERT gagal mendeteksi promosi yang menggunakan singkatan numerik.

b. DeBERTa menggunakan Byte-Pair Encoding (BPE) yang bekerja pada level byte dan memperlakukan setiap karakter termasuk spasi sebagai bagian dari token. BPE menggunakan penanda khusus "Ġ" untuk menunjukkan

adanya spasi sebelum kata. Pendekatan ini lebih robust terhadap variasi penulisan karena tidak bergantung pada kosakata tetap, melainkan membangun representasi dari unit-unit byte yang lebih fundamental. Namun, mekanisme disentangled attention DeBERTa mampu menangkap pola posisional dan kontekstual yang mengindikasikan makna "deposit" melalui pembelajaran dari contoh serupa dalam data training.

3. *Keterbatasan dan Tantangan Linguistik:* Meskipun performa kuantitatif kedua model sangat tinggi (> 99%), analisis linguistik mengungkapkan beberapa keterbatasan yang perlu menjadi perhatian, yaitu sensitivitas terhadap konteks negative dan ketergantungan pada kata kunci pada analisis word cloud.

Secara keseluruhan, analisis linguistik mengonfirmasi bahwa DeBERTa memiliki keunggulan signifikan dalam menangani ambiguitas bahasa dan variasi linguistik dibandingkan DistilBERT, terutama berkat mekanisme *disentangled attention* dan metode tokenisasi BPE yang lebih robust terhadap teks informal. Namun, tantangan dalam memahami negasi kontekstual dan adaptasi terhadap teknik baru masih menjadi area yang memerlukan pengembangan lebih lanjut.

IV. KESIMPULAN

Penelitian ini membandingkan performa model DistilBERT dan DeBERTa dalam mendeteksi promosi judi online pada komentar YouTube. Hasil evaluasi menunjukkan bahwa kedua model memiliki kinerja yang sangat tinggi dengan nilai presisi, recall, dan F1-score mencapai 1.00 (100%). DistilBERT memperoleh akurasi sebesar 99,29%, sedangkan DeBERTa sedikit lebih unggul dengan akurasi 99,84%. Analisis word cloud juga mengonfirmasi bahwa teks promosi judi online didominasi oleh kata-kata ajakan dan bonus seperti freebet dan gratis, sedangkan teks bukan promosi lebih bervariasi dan kontekstual. Oleh karena itu, DeBERTa layak dijadikan pilihan utama dalam membangun sistem deteksi otomatis konten ilegal di media sosial. Namun, DistilBERT tetap berguna apabila fokus utamanya adalah kecepatan inferensi dan efisiensi komputasi.

Dengan performa yang sangat baik tersebut, model deteksi ini berpotensi diterapkan sebagai sistem pendukung moderasi konten pada platform YouTube maupun sistem filter konten milik pemerintah. Integrasi model DeBERTa ke dalam mekanisme moderasi otomatis dapat membantu mendeteksi dan memblokir komentar promosi judi online secara real time, sehingga mampu meningkatkan keamanan digital dan menjaga kualitas interaksi pengguna. Ke depan, penelitian ini dapat dikembangkan dengan mengeksplorasi model *multilingual transformer* seperti IndoBERTweet atau mBERT agar mampu mengenali variasi bahasa informal dan campuran bahasa asing yang umum digunakan di media sosial. Selain itu, penerapan pendekatan *semi-supervised learning* juga dapat menjadi arah penelitian berikutnya untuk memanfaatkan data komentar tanpa label secara adaptif. Dengan demikian, sistem deteksi yang dihasilkan akan semakin tangguh, kontekstual, dan relevan dalam menghadapi dinamika penyebaran konten promosi ilegal di ruang digital.

DAFTAR PUSTAKA

[1] RRI Data (2024), Data Statistika Penggunaan Media Sosial Masyarakat Indonesia. https://www.rri.co.id/iptek/721570/ini-data-statistik-penggunaan-media-sosial-masyarakat-indonesia-tahun-2024.

[2] R. Bayu Perdana, I. Budi, A. Budi Santoso, A. Ramadiah, and P. Kresna Putra, "Detecting Online Gambling Promotions on Indonesian Twitter Using Text Mining Algorithm," 2024. [Online]. Available: www.ijacsa.thesai.org

[3] A. Maulana and A. Yuliana, "Analisis Sentimen Opini Publik Terkait Judi Online Pada Pengguna Aplikasi X Menggunakan Algoritma Naïve Bayes Dan Support Vector Mechine," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5187.

[4] E. Smith, N. Reiter, and J. Peters, "Automatic detection of problem-gambling signs from online texts using large language models," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2312.00804

[5] F. Fajri *et al.*, "Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter," vol. 8, no. 2, pp. 71–80, 2022.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[7] J. Carreras Timoneda and S. Vallejo Vera, "BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text," *J Polit*, Jan. 2024, doi: 10.1086/730737.

[8] S. Ruder, M. Peters, S. Swayamdipta, and T. Wolf, "Ruder, S. (2019). Transfer Learning in Natural Language Processing. arXiv preprint arXiv:1901.11504."

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[10] A. Chopra, A. Prashar, and C. Sain, "Natural Language Processing," *International Journal Of Technology Enhancements And Emerging Engineering Research*, vol. 1, no. 4, 2013, [Online]. Available: http://en.wikipedia.org/wiki/

[11] C. Jocelynne, L. Tobing, I. Lanang Wijayakusuma, L. Putu, and I. Harini, "Detection of Political Hoax News Using Fine-Tuning IndoBERT," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[12] Vasmani, A., Shazeer, N., Parman, N., Uszkoreit, J., Jones, L., Gomez, A., & Kaiser, L. (2017) Attention is all you need. URL https://proceedings.neurips.cc/paper_files/paper/ 2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf."

[13] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Jun. 2020, doi: https://doi.org/10.48550/arXiv.2006.03654.

[14] M. B. Nugroho, A. Khanif Zyen, and A. Widiastuti, "Multiclass Sentiment Analysis of Electric Vehicle Incentive Policies Using IndoBERT and DeBERTa Algorithms," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] S. Mahendru and T. Pandit, "SecureNet: A Comparative Study of DeBERTa and Large Language Models for Phishing Detection," Jun. 2024, doi: 10.1109/BDAI62182.2024.10692765.