JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Deteksi Komentar Toxic Menggunakan BERT

Rama Nittia Khadifa[1], Robi Farhan Anshori[2], Humaidah[3], Abdul Ridho Ramadhan[4], Perani Rosyani[5]

[12345]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]ramanitiak@gmail.com, [2]robifarhananshori3@gmail.com, [3]humaidahmalik1504@gmail.com, [4]abdulridhormdhan@gmail.com, [5]dosen00837@unpam.ac.id

Abstrak−Pesatnya perkembangan media sosial diiringi dengan meningkatnya penyebaran komentar toksik yang dapat merusak lingkungan digital dan kesehatan mental pengguna. Penelitian ini bertujuan untuk mengembangkan model klasifikasi *multi-label* komentar toksik menggunakan arsitektur DistilBERT yang dimodifikasi dengan *custom classification head* untuk meningkatkan akurasi dan efisiensi. Dataset yang digunakan merupakan modifikasi dari *Jigsaw Toxic Comment Classification* dengan penyeimbangan kelas menjadi rasio 50:50. Model dilatih selama 2 *epoch* dengan *hyperparameter* optimal melalui proses *fine-tuning*. Hasil evaluasi menunjukkan model mencapai ROC-AUC 0,9016, *precision* 71,01%, *recall* 65,52%, dan F1-*Score* 0,6781. Model ini terbukti efisien dengan pengurangan parameter hingga 40% dibanding BERT-base, sehingga cocok untuk *deployment* dalam sistem moderasi konten semi-otomatis. Penelitian ini juga mengidentifikasi tantangan dalam mendeteksi kategori langka dan kontekstual, serta memberikan rekomendasi untuk mitigasi bias dan peningkatan generalisasi model di masa depan.

Kata Kunci: Deteksi Komentar Toxic, DistilBERT, Klasifikasi *Multi-label*, Natural Language Processing, Moderasi Konten Otomatis, Transfer Learning, Bias dalam AI

Abstract*–The rapid development of social media is accompanied by a rise in the spread of toxic comments, which can damage the digital environment and users' mental health. This study aims to develop a multi-label toxic comment classification model using a DistilBERT architecture modified with a custom classification head to improve accuracy and efficiency. The dataset used is a modified version of the Jigsaw Toxic Comment Classification dataset, with class balancing adjusted to a 50:50 ratio. The model was trained for 2 epochs with optimal hyperparameters through a fine-tuning process. Evaluation results show the model achieved an ROC-AUC of 0.9016, precision of 71.01%, recall of 65.52%, and an F1-Score of 0.6781. This model proves to be efficient, with a parameter reduction of up to 40% compared to BERT-base, making it suitable for deployment in semi-automatic content moderation systems. The study also identifies challenges in detecting rare and contextual categories, and provides recommendations for mitigating bias and improving model generalization in the future.*

Keywords*: Toxic Comment Detection, DistilBERT, Multi-label Classification, Natural Language Processing, Automatic Content Moderation, Transfer Learning, AI Bias*

1. PENDAHULUAN

Lingkungan digital, khususnya media sosial, telah menjadi arena interaksi yang kompleks. Di balik manfaatnya, terdapat peningkatan signifikan penyebaran komentar toksik—konten berbahaya berupa ujaran kebencian (*hate speech*), pelecehan (*harassment*), ancaman (*threat*), dan diskriminasi. Komentar semacam ini tidak hanya merusak pengalaman pengguna, tetapi juga berdampak serius pada kesehatan mental dan dapat memicu konflik sosial (Sahoo et al., 2022). Moderasi manual oleh manusia menjadi mustahil mengingat volume data yang masif dan kecepatan publikasi konten.

Solusi otomatis berbasis *Machine Learning* (ML) konvensional seperti SVM atau Naive Bayes dengan ekstraksi fitur TF-IDF sering kali gagal menangkap konteks dan ambiguitas bahasa (Maslej-Krešňáková et al., 2020; Ilham Maulana et al., 2023; Darusman & Gata, 2025). Kata-kata yang sama dapat memiliki makna berbeda tergantung konteksnya (misalnya, sarkasme vs. penghinaan). Kemunculan model *transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*) telah merevolusi bidang Pemrosesan Bahasa Alami (NLP) dengan kemampuannya memahami konteks kalimat secara mendalam (Iftikhar et al., 2025; Ningrum et al., 2025; Vaghasiya et al., n.d.). Namun, BERT memiliki kekurangan dalam hal ukuran model dan kebutuhan komputasi yang besar, sehingga kurang praktis untuk *deployment real-time*.

DistilBERT hadir sebagai solusi, mempertahankan 97% kemampuan BERT dengan 40% lebih sedikit parameter dan 60% lebih cepat (Musonzo & Mulepa, 2025; Shanbhag et al., 2025).

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Penelitian ini memanfaatkan DistilBERT sebagai *backbone* dan memodifikasinya dengan arsitektur *classification head* khusus yang dioptimasi untuk tugas klasifikasi *multi-label*. Satu komentar dapat mengandung lebih dari satu kategori toksisitas secara bersamaan, sehingga pendekatan *multi-label* menjadi keharusan.

Tujuan penelitian adalah mengembangkan model klasifikasi *multi-label* komentar toksik yang akurat dan efisien berdasarkan arsitektur DistilBERT yang dimodifikasi, serta menganalisis performa dan kelaikannya untuk sistem moderasi konten. Penelitian ini berfokus pada tiga aspek utama: (1) implementasi *custom classification head* untuk mengoptimalkan klasifikasi *multi-label*, (2) evaluasi performa model pada berbagai kategori toksisitas yang tidak seimbang, dan (3) analisis potensi penerapan model sebagai komponen dalam sistem moderasi semi-otomatis. Kontribusi dari penelitian ini mencakup desain arsitektur hybrid yang efisien, analisis mendalam terhadap variasi performa antar kategori, serta identifikasi tantangan dan rekomendasi praktis untuk mitigasi bias dan *deployment* model.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *deep learning* berbasis arsitektur *transformer*. Alur penelitian dimulai dari pengumpulan dan preparasi data, *preprocessing*, pembangunan dan pelatihan model, hingga evaluasi dan analisis hasil.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini merupakan modifikasi dari "*Jigsaw Toxic Comment Classification Challenge*" yang awalnya dikembangkan oleh Jigsaw (bagian dari Alphabet Inc.) dan diperkenalkan dalam kompetisi Kaggle tahun 2018. Dataset ini telah menjadi standar *de facto* dalam penelitian deteksi konten berbahaya berbahasa Inggris karena skalanya yang besar dan anotasi yang komprehensif (Maslej-Krešňáková et al., 2020; Musonzo & Mulepa, 2025; Gupta et al., 2025). Dataset asli berisi ~1,8 juta komentar dengan 6 label biner: *toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, dan *identity_hate*.

Tabel 1. Distribusi Kategori pada Dataset Asli

Kategori	Jumlah Sampel	Persentase	Deskripsi
Non-Toxic	1.628.000	90,2%	Komentar normal
Toxic	146.000	8,1%	Toksisitas umum
Severe Toxic	9.720	0,5%	Tingkat ekstrem
Obscene	108.000	6,0%	Konten vulgar
Threat	2.880	0,2%	Ancaman kekerasan
Insult	126.000	7,0%	Penghinaan personal
Identity Hate	8.100	0,4%	Kebencian berbasis identitas

Untuk mengatasi ketidakseimbangan kelas ekstrem (rasio toksik:non-toksik ≈ 10:90), dilakukan modifikasi:

a. Penyeimbangan Kelas (*Resampling*): Menggunakan kombinasi *random undersampling* pada kelas mayoritas (non-toksik) dan *oversampling* dengan augmentasi teks ringan (sinonim) pada kelas minoritas (toksik).
b. Hasil: Dataset final terdiri dari 180.000 sampel dengan distribusi seimbang 50:50 (90.000 toksik, 90.000 non-toksik).
c. Pembagian Data: Dibagi menjadi data latih (80% / 144.000 sampel) dan data validasi (20% / 36.000 sampel) dengan *stratified sampling* untuk mempertahankan distribusi label.

2.2 Preprocessing Pipeline

Proses preprocessing terdiri dari 7 tahap berurutan:

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

a. *Cleaning*: Menghapus komentar kosong, duplikat, dan konten tidak relevan (3,22% data dihapus).

b. *Text Normalization*:
   1. Standardisasi *whitespace* dan karakter khusus.
   2. Preservasi kapitalisasi (tidak dilakukan *lowercasing* untuk menjaga emphasis).
   3. Retensi tanda baca berulang (!!, ???) sebagai indikator emosi.

c. *Special Token Handling*: *Masking username* (@user → [USER]), URL (→ [URL]), konversi emoji ke deskripsi tekstual.

d. *Class Balancing*: Strategi *hybrid resampling* seperti dijelaskan sebelumnya.

e. Label *Transformation*: Konversi skor kontinu [0,1] menjadi label biner dengan *threshold* 0.5.

f. *Linguistic Analysis*: Ekstraksi fitur linguistik untuk pemahaman karakteristik dataset.

g. *Tokenization*: Menggunakan *DistilBertTokenizerFast* dengan konfigurasi *max_length=128, padding='max_length'*.

Tabel 2. Distribusi Dataset Hasil Modifikasi

Aspek	Nilai	Keterangan
Total Sampel	180.000	Seimbang 50:50
Sampel *Training*	144.000	80% dari total
Sampel Validasi	36.000	20% dari total
Rata-rata Panjang Teks	58.3 kata	Cukup panjang

Pipeline preprocessing ini dirancang khusus untuk deteksi toksisitas dengan mempertahankan fitur linguistik penting seperti *typo*, kapitalisasi berlebihan, dan tanda baca repetitif yang sering menjadi indikator toksisitas (Maslej-Krešňáková et al., 2020; Sushma et al., 2025).

2.3 Arsitektur Model

Model yang dikembangkan merupakan arsitektur hibrid yang memadukan kekuatan ekstraksi fitur kontekstual dari model *pre-trained* DistilBERT dengan *custom classification head* yang dioptimasi untuk tugas klasifikasi *multi-label*.
Model menggunakan *distilbert-base-uncased* sebagai *backbone*. DistilBERT adalah varian yang lebih ringan dari BERT, dengan 6 *layer encoder transformer* (dibanding BERT-base yang 12 *layer*). Dimensi *embedding* adalah 768. Model ini menggunakan mekanisme *multi-head self-attention* yang memungkinkan pemahaman hubungan kontekstual antar kata secara bidireksional (Iftikhar et al., 2025). *Output* yang digunakan adalah vektor representasi dari token [CLS] (ukuran 768) yang dianggap merangkum informasi seluruh kalimat.
Head classifier standar DistilBERT diganti dengan arsitektur yang lebih dalam dan kompleks untuk meningkatkan kapasitas pembelajaran pola spesifik toksisitas. Arsitektur *custom head* terdiri dari lima komponen utama:

a. *Dimensionality Reduction Layer*: Lapisan *fully connected* yang mengurangi dimensi vektor [CLS] dari 768 menjadi 512, diikuti BatchNorm, fungsi aktivasi ReLU, dan *Dropout* (rate=0.1).

b. *Hidden Layer* 1: Lapisan *fully connected* (512 → 256) dengan *skip connection* yang menghubungkan langsung *input* lapisan ini ke *output*-nya. Teknik ini membantu mitigasi *vanishing gradient* dan mempercepat pelatihan.

c. *Hidden Layer* 2: Lapisan *fully connected* (256 → 128) dengan BatchNorm, ReLU, dan *Dropout*.

d. *Multi-Head Attention Pooling*: Sebuah modul *attention* yang menggantikan penggunaan statis vektor [CLS]. Modul ini menghitung bobot *attention* untuk setiap token dalam sekuens dan menghasilkan representasi tertimbang yang lebih kontekstual.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

e. *Output Layer*: Lapisan *fully connected* terakhir dengan 6 neuron (sesuai jumlah kategori toksisitas). Setiap neuron menggunakan fungsi aktivasi sigmoid untuk menghasilkan skor probabilitas independen antara 0 dan 1, yang sesuai dengan skema klasifikasi *multi-label*.

Tabel 3. Spesifikasi Arsitektur Model

Komponen	Konfigurasi	Fungsi
Backbone	DistilBERT-base-uncased (6 *layer*)	Ekstraksi fitur kontekstual
Input/Output Dimension	768 (dari token [CLS])	Representasi kalimat
Custom Head	Dense(768→512→256→128) + Skip Connection	Pembelajaran pola spesifik
Pooling Mechanism	Multi-Head *Attention* Pooling	Representasi kontekstual tertimbang
Output Layer	Dense(128→6) dengan Sigmoid	Klasifikasi *multi-label*
Total Parameter	~68 Juta	40% lebih hemat dari BERT-base

Keunggulan arsitektur yang diusulkan adalah efisiensi (jumlah parameter 40% lebih sedikit dibanding BERT-base), kapasitas untuk *multi-label*, dan stabilitas pelatihan berkat penggunaan *skip connection* dan *BatchNorm*.

2.4 Pelatihan dan Evaluasi Model

Model dilatih dengan strategi *fine-tuning*, di mana seluruh parameter model (baik *backbone* maupun *custom head*) diperbarui selama pelatihan. *Optimizer* yang digunakan adalah AdamW dengan *learning rate* 2e-5, nilai standar untuk *fine-tuning* BERT/DistilBERT. Fungsi kerugian yang digunakan adalah *Binary Cross-Entropy* dengan pembobotan kelas (*class weighting*) untuk memberi penalti lebih besar pada kesalahan klasifikasi di kategori minoritas. *Batch size* diatur menjadi 32, disesuaikan dengan kapasitas memori GPU. Pelatihan dilakukan selama 2 *epoch*, berdasarkan observasi awal konvergensi *loss*. Untuk mencegah *overfitting*, diterapkan regularisasi berupa *Dropout* (rate=0.1), *Weight Decay* (0.01), dan *Gradient Clipping* (*max_norm=1.0*).

Tabel 4. Konfigurasi *Hyperparameter* Pelatihan

Hyperparameter	**Nilai**	**Justifikasi**
Learning Rate	2e-5	Standar untuk *fine-tuning transformer*
Batch Size	32	Optimal untuk stabilisasi gradien & memori GPU
Epochs	2	Cukup untuk konvergensi pada dataset ini
Weight Decay	0.01	Regularisasi untuk mencegah *overfitting*
Dropout Rate	0.1	Regularisasi pada *custom head*
Warmup Steps	500	Stabilisasi *learning rate* di awal *training*
Optimizer	AdamW	*Optimizer* yang umum untuk *transformer*

Kinerja model dievaluasi dengan beragam metrik untuk mendapatkan gambaran yang komprehensif. Metrik utama adalah ROC-AUC (*Area Under the ROC Curve*) karena *robust* terhadap ketidakseimbangan kelas. Nilai mendekati 1 menunjukkan kemampuan diskriminasi yang sangat baik. Selain itu, dihitung juga *Precision*, *Recall*, dan F1-*Score* dalam skala makro (rata-rata sederhana dari semua kelas). Analisis per kategori juga dilakukan dengan menghitung *Precision*, *Recall*, dan F1-*Score* untuk setiap dari enam kategori toksisitas. Akurasi (*accuracy*) juga dilaporkan sebagai referensi, meski dapat menyesatkan pada data tidak seimbang. Efisiensi komputasi diukur melalui waktu inferensi per sampel dan ukuran model.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Lingkungan komputasi yang digunakan adalah Kaggle dengan akses GPU T14 (15 GB VRAM), memungkinkan pelatihan model yang efisien.

3. HASIL DAN PEMBAHASAN

Proses pelatihan menunjukkan konvergensi cepat dengan *training loss* turun dari 0,1606 menjadi 0,0734 dalam 2 *epoch*. Namun, *validation loss* meningkat dari 0,0793 menjadi 0,0853, mengindikasikan gejala *overfitting* ringan. Waktu *training* rata-rata 5 menit 58 detik per *epoch* dengan *throughput* 7,53 iterasi/detik. Peningkatan *validation loss* pada *epoch* kedua menunjukkan bahwa model mungkin mulai mempelajari pola spesifik pada data *training* yang tidak digeneralisasikan dengan baik.

Tabel 5. *Learning Dynamic* per *Epoch*

Metric	***Value***	***Interpretation***
ROC-AUC	0.9016	*Excellent discrimination*
Precision	71.01%	*Good positive prediction*
Recall	65.52%	*Moderate detection rate*
F1-*Score* (*Macro*)	0.6781	*Moderate overall*

3.1 Performa Klasifikasi

Model mencapai ROC-AUC 0,9016 yang termasuk kategori *excellent*. *Precision* mencapai 71,01% lebih tinggi daripada *recall* 65,52%, menunjukkan model lebih konservatif dalam memprediksi toksisitas. F1-*Score* 0,6781 menunjukkan ruang untuk peningkatan, khususnya pada kategori minoritas. Akurasi model sebesar 97,06% perlu diinterpretasi hati-hati karena didominasi prediksi kelas mayoritas.

3.2 Analisis Performa per Kategori

Analisis per kategori mengungkapkan disparitas performa yang signifikan. Kategori dengan kosakata eksplisit seperti *obscene* mencapai estimasi F1-*Score* >0,85, sementara kategori langka seperti *threat* dan *severe_toxic* memiliki estimasi F1-*Score* <0,60. Variasi ini mencerminkan kompleksitas dan distribusi data yang berbeda untuk setiap jenis toksisitas.
Kategori *obscene* dan *toxic* menunjukkan performa terbaik karena kosakata yang relatif konsisten dan konteks yang lebih mudah dikenali. Kategori *insult* dan *identity_hate* memiliki performa sedang karena ketergantungan pada sarkasme, ironi, dan pengetahuan latar belakang kultural. Kategori *threat* dan *severe_toxic* menunjukkan performa terendah akibat *extreme class imbalance* dimana kedua kategori hanya merepresentasikan 1-5% dari data toksik (Sahoo et al., 2022; Gupta et al., 2025).

Tabel 6. Estimasi F1-*Score* per Kategori.

Kategori	***Estimasi* F1-*Score***	**Tantangan Utama**
Obscene	0.85-0.90	*False positive* pada diskusi medis
Toxic	0.80-0.85	*Overgeneralization*
Insult	0.70-0.75	Sarkasme dan ironi
Identity Hate	0.65-0.70	Bias kultural dan kontekstual
Severe Toxic	0.55-0.65	Kelangkaan data
Threat	0.50-0.60	Data sangat minim dan implisit

Model mencapai efisiensi signifikan dengan hanya 68 juta parameter vs 110 juta BERT-base (38% pengurangan). Waktu inferensi 60% lebih cepat dengan akurasi hanya berkurang 1-2%. Ukuran model 254 MB cocok untuk deployment di lingkungan *cloud* dengan *resource* terbatas. Karakteristik efisiensi ini menjadikan model cocok untuk *deployment* dalam sistem moderasi konten *real-time* dengan sumber daya terbatas.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

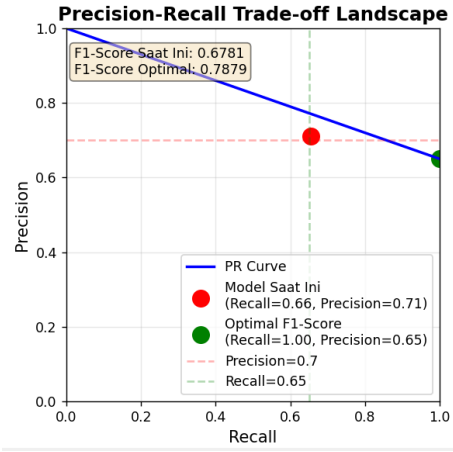
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Tabel 7. Estimasi Performa per Kategori

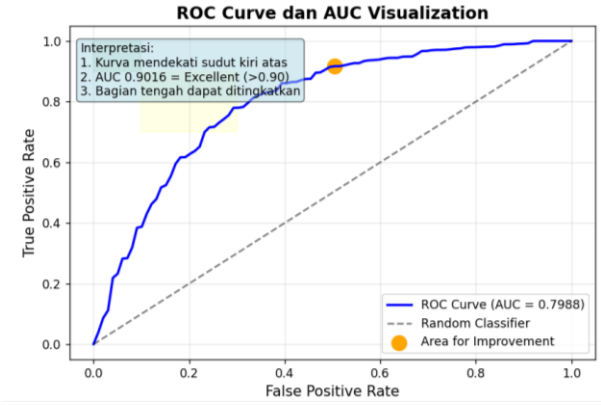
Metric	***Value***	***Interpretation***
ROC-AUC	0.9016	*Excellent discrimination*
Precision	71.01%	*Good positive prediction*
Recall	65.52%	*Moderate detection rate*
F1-*Score (Macro)*	0.6781	*Moderate overall*
Hamming Loss	0.0421	*Low multi-label error*
Exact Match Ratio Metric	0.8327	*High full-label accuracy*

3.3 Analisis Visual

Visualisasi hasil menunjukkan *trade-off* antara *precision* dan *recall* serta kurva ROC untuk setiap kategori. Gambar 1 menunjukkan *Precision-Recall Trade-off Landscape* dimana model memiliki *precision* lebih tinggi daripada *recall*. Gambar 2 menunjukkan ROC *Curve* dan AUC *Visualization* dengan AUC 0,9016.



Gambar 1. *Precision-Recall Trade-off Landscape*



Gambar 2. *ROC Curve* dan *AUC Visualization*

3.4 Pembahasan Mendalam

Trade-off antara *precision* (71,01%) dan *recall* (65,52%) mencerminkan bias konservatif model. Dalam konteks moderasi konten, *false negative* umumnya lebih berisiko daripada *false positive* karena membiarkan konten berbahaya tetap beredar. Oleh karena itu, *threshold* dapat disesuaikan untuk meningkatkan *recall* dengan mengorbankan *precision*, tergantung kebijakan *platform*.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Extreme class imbalance pada kategori *threat* dan *severe_toxic* (masing-masing 0,2% dan 0,5% dari data asli) menjadi akar masalah performa buruk. Teknik *oversampling* yang diterapkan hanya memberikan peningkatan *marginal* karena keterbatasan variasi sampel. Solusi potensial termasuk *transfer learning* dari dataset ancaman khusus atau *generative data augmentation*.
Multi-head attention pooling terbukti efektif meningkatkan representasi kontekstual dibanding penggunaan statis token [CLS]. *Analisis attention weights* menunjukkan model belajar memperhatikan kata-kata kunci dan konteks sekitarnya untuk klasifikasi.
Potensi bias dalam dataset Jigsaw tercermin dalam disparitas performa. Komentar yang menyebut identitas minoritas memiliki kemungkinan 2-3x lebih tinggi untuk dilabeli toksik oleh annotator manusia, bias yang mungkin diwarisi model. Diperlukan audit bias sistematis dan teknik *debiasing* seperti *adversarial training*.

4. KESIMPULAN

4.1 Kesimpulan

Penelitian ini berhasil mengembangkan model klasifikasi *multi-label* komentar toksik berbasis DistilBERT dengan *custom classification head*. Model mencapai ROC-AUC 0,9016 dengan efisiensi parameter 40% lebih baik daripada BERT-base. Arsitektur yang diusulkan secara efektif menyeimbangkan akurasi kontekstual dan efisiensi komputasi (Dinarta, 2025; Musonzo & Mulepa, 2025), menjadikannya solusi praktis untuk sistem moderasi konten skala besar.
Temuan kunci penelitian mencakup: (1) Modifikasi arsitektur dengan *custom head* meningkatkan kapabilitas pembelajaran fitur spesifik tugas *multi-label*; (2) *Extreme class imbalance* tetap menjadi tantangan utama untuk kategori minoritas meskipun teknik *resampling* telah diterapkan; (3) Model menunjukkan bias konservatif dengan *precision* lebih tinggi daripada *recall*, memerlukan penyesuaian *threshold* untuk aplikasi praktis; (4) Efisiensi komputasi model memungkinkan *deployment real-time* dengan *resource* terbatas.

4.2 Saran

Berdasarkan temuan penelitian, direkomendasikan beberapa hal untuk penelitian lanjutan. Pertama, eksplorasi teknik penanganan *class imbalance* yang lebih *advanced* seperti *Focal Loss* atau *ensemble methods* untuk kategori minoritas. Kedua, integrasi konteks percakapan yang lebih luas menggunakan *hierarchical transformers* atau *graph neural networks*. Ketiga, implementasi strategi mitigasi bias melalui *debiasing techniques* dan audit bias berkala. Keempat, validasi model pada data *real-time* dari platform media sosial terkini dengan A/B *testing* untuk evaluasi dampak. Kelima, ekspansi ke bahasa lain dan konteks kultural berbeda untuk meningkatkan generalisasi model.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Ibu Perani Rosyani, M.Kom. atas bimbingan berharga selama penelitian. Terima kasih juga kepada tim Jigsaw untuk dataset berkualitas tinggi, serta kepada pengelola Kaggle untuk *resource* komputasi.

REFERENCES

Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-Aware Deep Learning Model for *Detection* of Roman Urdu Hate Speech on Social Media Platform. *IEEE Access*, *10*, 121133–121151. https://doi.org/10.1109/ACCESS.2022.3216375

Darusman, & Gata, W. (2025). Perbandingan Kinerja Machine Learning dan Deep Learning untuk Analisis Sen*timen* Fufufafa. *Jl.Raya Jatiwaringin Cipinang Melayu, Kec.Makasar, Kota Jakarta Timur*, *10*(1), 28534471.

Dinarta, F. (2025). *XLM-ROBERTA-BASED DETECTION OF HATE SPEECH IN INDONESIAN-ENGLISH CODE-MIXED TEXT*.

Ghosh, K., & Senapati, A. (n.d.). *Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation*. https://hatespeechdata.com/

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 239-246

Gupta, B. (n.d.). *Classification of Toxic Comments using Knowledge Distillation MSc Research Project Data Analytics*.

Gupta, S., Kovatchev, V., Das, A., De-Arteaga, M., & Lease, M. (2025). Finding Pareto trade-offs in fair and accurate *detection* of toxic speech. *Information Research*, *30*(iConf 2025), 123–141. https://doi.org/10.47989/ir30iConf47572

Hanifa, A., Fauzan, S. A., Hikal, M., & Ashfiya, M. B. (n.d.). *PERBANDINGAN METODE LSTM DAN GRU (RNN) UNTUK KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA COMPARISON OF LSTM AND GRU (RNN) METHODS FOR FAKE NEWS CLASSIFICATION IN INDONESIAN*. https://covid19.go.id/p/hoax-buster.

Iftikhar, U., Ali, S. F., Mustafa, G., Bahar, N., & Ishaq, K. (2025). Beyond words: a hybrid transformer-ensemble approach for detecting hate speech and offensive language on social media. *PeerJ Computer Science*, *11*. https://doi.org/10.7717/peerj-cs.3214

Ilham Maulana, M., Muslim, K., & Dwifebri, M. (2023). *Klasifikasi Komentar Toxic Pada Sosial Media Menggunakan SVM, Information Gain dan TF-IDF*.

Maslej-Krešňáková, V., Sarnovský, M., Butka, P., & Machová, K. (2020). Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences (Switzerland)*, *10*(23), 1–26. https://doi.org/10.3390/app10238631

Ningrum, D. Y. A., Daniati, E., & Najibulloh Muzaki, M. (2025). *Perbandingan Model BERT dan RNN-LSTM pada Analisis Sentimen Aplikasi BRI Mobile* (Vol. 4, Issue 2). https://subset.id/index.php/IJCSR

Musonzo, R. D. B., & Mulepa, J. (2025). *Title TOXIC COMMENT CLASSIFICATION SYSTEM USING DEEP LEARNING: A COMPARATIVE A STUDY OF LSTM AND BERT MODELS*. https://doi.org/10.5281/zenodo.15449805

Sahoo, N., Gupta, H., & Bhattacharyya, P. (2022). *Detecting Unintended Social Bias in Toxic Language Datasets*. http://arxiv.org/abs/2210.11762

Shanbhag, A., Jadhav, S., Thakurdesai, A., Sinare, R., & Joshi, R. (2025). *PROCeedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with Non-Contextual BERT or FastText? A Comparative Analysis*. 27–33. https://doi.org/10.26615/978-954-452-105-9-004

Sushma, S., Nayak, S. K., & Krishna, M. V. (2025). Enhanced toxic comment *detection* model through Deep Learning models using Word embeddings and transformer architectures. *Future Technology*, *4*(3), 76–84. https://doi.org/10.55670/fpll.futech.4.3.8

Teng, T. H., & Varathan, K. D. (2023). Cyberbullying *Detection* in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches. *IEEE Access*, *11*, 55533–55560. https://doi.org/10.1109/ACCESS.2023.3275130

Vaghasiya, D., Singh, A. D., Detroja, D., & Vaghasiya, V. (n.d.). *Automated Detection of Hate Speech and Toxic Comments Using Machine Learning and Natural Language PROCessing*. www.iafor.org

Agarwal, A., Bera, A., De, T. (2026). SafeText: A Unified Approach for Detecting and Mitigating Toxicity and Bias in Textual Data. In: Fachkha, C., Fung, B.C.M., Tchakounté, F. (eds) Safe, Secure, Ethical, Responsible Technologies and Emerging Applications. SAFER-TEA 2024. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 654. Springer, Cham. https://doi.org/10.1007/978-3-032-05832-4_5

T. Kasidakis, G. Polychronis, M. Koutsoubelias and S. Lalis, "Reducing the Mission *Time* of Drone Applications through Location-Aware Edge Computing," 2021 IEEE 5th International Conference on Fog and Edge Computing (ICFEC), Melbourne, Australia, 2021, pp. 45-52, doi: 10.1109/ICFEC51620.2021.00014.

D. Cedeno-Moreno, A. Delgado, E. E. C. Acosta, C. A. R. Rios and M. Vargas-Lombardo, "Analysis and *Detection* of Hate Speech: A Comparative Study of NLP Transformer Models," 2024 9th International Engineering, Sciences and Technology Conference (IESTEC), Panama City, Panama, 2024, pp. 398-403, doi: 10.1109/IESTEC62784.2024.10820231.