Analisis Performansi *Hate Comments* pada *Learning Rate* 10^{-1}- 10^{-3} dengan Dataset dari X

1st Anggara Budiyanto
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
anggarabudiyanto@student.telkomuniversity.ac.id

2nd Suryo Adhi Wibowo
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
suryoadhiwibowo@telkomuniversity.ac.id

3rd Korediantо Usman
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
korediantousman@telkomuniversity.ac.id

***Abstrak*—*Cyberbullying* merupakan fenomena sosial yang semakin meningkat seiring dengan meningkatnya penggunaan media sosial, dan seringkali menyebabkan dampak psikologis serta emosional yang merugikan, terutama melalui *hate comments*. Penelitian ini bertujuan untuk mengevaluasi kinerja model IndoBERT dan Cendol dalam mendeteksi komentar kebencian yang berhubungan dengan *cyberbullying*. Survei terhadap 328 partisipan menghasilkan 64 kata kunci terkait *cyberbullying*. Proses penelitian mencakup pengumpulan dataset yang berisi kata kunci tersebut, serta pengujian kedua model menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-Score*. Hasil eksperimen menunjukkan bahwa model Cendol unggul dengan akurasi sebesar 90,5% pada konfigurasi *batch size* 15, *epoch* ke-4, dan *learning rate* 10^{-3}, sementara IndoBERT hanya mencapai akurasi 36% pada konfigurasi *batch size* 5, *epoch* ke-4, dan *learning rate* 10^{-3}. Meskipun kedua model menunjukkan potensi dalam mendeteksi ujaran kebencian, model IndoBERT menunjukkan performa yang lebih rendah pada dataset yang digunakan, kemungkinan disebabkan oleh keterbatasan dalam menangani konteks lokal. Penelitian ini memberikan kontribusi signifikan dalam pengembangan teknologi deteksi ujaran kebencian berbasis bahasa Indonesia, yang dapat diimplementasikan pada berbagai platform media sosial seperti X, Facebook, Instagram, dan TikTok untuk mengurangi dampak negatif dari *hate comments*.**

***Kata Kunci*: *Cyberbullying*, *Hate Comments*, IndoBERT, Cendol, NLP.**

I. PENDAHULUAN

Cyberbullying dan ujaran kebencian telah menjadi masalah sosial yang serius di era digital. Kemajuan teknologi komunikasi memungkinkan interaksi lintas batas, namun juga membuka peluang terjadinya tindakan intimidasi dan pelecehan secara daring. Fenomena ini mencakup berbagai bentuk, mulai dari komentar negatif yang menyerang aspek pribadi seperti ras, agama, atau orientasi seksual, hingga ancaman langsung tanpa alasan yang jelas. Contoh nyata meliputi kasus perundungan di lingkungan pendidikan dan penyebaran ujaran kebencian dalam konteks politik, seperti yang terjadi selama kampanye Pemilu 2024 [1]. Media sosial telah menjadi *platform* utama yang memfasilitasi penyebaran tindakan tersebut, dengan jumlah unggahan berisi ujaran kebencian di Indonesia yang terus bertambah.

Sebagai salah satu wujud perundungan digital, *hate comments* (komentar kebencian) memiliki dampak psikologis yang signifikan bagi korban. Tidak hanya menyulut permusuhan, tindakan ini juga memperburuk ketegangan sosial. Berbagai studi mengungkapkan bahwa *platform* media sosial seperti X, Facebook, dan Instagram sering menjadi sarana utama penyebaran ujaran kebencian. Sebagai contoh, selama masa kampanye Pemilu 2024, *platform* X mencatat lebih dari 120 ribu unggahan dengan konten ujaran kebencian, menjadikannya *platform* dengan kontribusi terbesar terhadap masalah tersebut [2]. Data ini menunjukkan betapa rumitnya tantangan dalam mengatasi *cyberbullying* dan ujaran kebencian secara

efektif.

Oleh karena itu, analisis kinerja model berbasis kecerdasan buatan seperti IndoBERT dan Cendol menjadi sangat penting. Model-model ini dirancang untuk mendeteksi dan mengklasifikasikan ujaran kebencian secara akurat. Penelitian tentang kemampuan IndoBERT dan Cendol dalam mendeteksi *hate comments* bertujuan tidak hanya untuk meningkatkan efektivitas deteksi, tetapi juga untuk menawarkan solusi praktis dalam meminimalkan dampak negatif *cyberbullying*. Dengan menganalisis pola dan karakteristik ujaran kebencian melalui data, pendekatan berbasis teknologi diharapkan dapat menjadi alat yang andal untuk menciptakan lingkungan digital yang lebih aman.

II. KAJIAN TEORI

A. Bidirectional Encoder Representations from Transformers (BERT)

BERT adalah model pembelajaran mesin yang dikembangkan oleh Google untuk menangani berbagai tugas dalam *Natural Language Processing* (NLP). Dengan menggunakan arsitektur *transformer*, BERT memungkinkan analisis konteks secara *bidirectional*, yaitu mempertimbangkan kata-kata di kedua sisi kata target dalam sebuah kalimat. Beberapa karakteristik utama dari BERT adalah:

- *Bidirectional*: Memahami konteks dari kedua arah untuk meningkatkan pemahaman kalimat secara keseluruhan.
- *Pre-training dan Fine-tuning*: Model ini dilatih menggunakan dataset besar (*pre-training*) sebelum diadaptasi untuk tugas tertentu (*fine-tuning*).
- Kemampuan *Transfer Learning*: BERT dapat diterapkan pada berbagai tugas NLP dengan penyesuaian minimal.

BERT sering digunakan untuk berbagai aplikasi NLP, termasuk deteksi ujaran kebencian. Sebagai contoh, setelah dilakukan *fine-tuning* menggunakan *optimizer* seperti AdamW, BERT mampu mendeteksi ujaran kebencian di Twitter dengan tingkat akurasi sekitar 90% [3].

BERT memiliki beberapa versi, di antaranya:

- *BERT-Base*: Terdiri dari 12 lapisan, 768 dimensi tersembunyi, 12 kepala perhatian, dan memiliki 110 juta parameter.
- *BERT-Large*: Memiliki 24 lapisan, 1024 dimensi tersembunyi, 16 kepala perhatian, dan 340 juta parameter.
- Varian kecil seperti *BERT-Tiny*, *BERT-Mini*, *BERT-Small*, dan *BERT-Medium*.
- *Multilingual BERT*, yang mendukung lebih dari 100 bahasa.

Parameter utama BERT meliputi ukuran kosakata, dimensi tersembunyi, jumlah lapisan transformer, dan jumlah kepala perhatian. Model ini dilatih menggunakan korpus besar berbahasa Inggris dan memiliki versi multibahasa, seperti *mBERT*, yang mendukung bahasa Indonesia, serta *IndoBERT*, yang dirancang khusus untuk bahasa Indonesia.

B. Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT)

IndoBERT adalah adaptasi dari model BERT yang dikembangkan khusus untuk bahasa Indonesia. Model ini dilatih menggunakan dataset teks besar dalam bahasa Indonesia, sehingga mampu menangkap konteks dan nuansa bahasa lokal dengan lebih baik. Salah satu varian yang digunakan dalam penelitian ini adalah *IndoBERT-Base*, dengan spesifikasi berikut:

- *Jumlah lapisan (L)*: 12
- *Ukuran tersembunyi (H)*: 768
- *Jumlah kepala perhatian (A)*: 12
- *Total parameter*: sekitar 125 juta
- *Pelatihan*: Dilatih menggunakan 5,5 miliar kata dari sumber seperti artikel berita dan Wikipedia.

IndoBERT-Base unggul dalam menangani berbagai tugas NLP, termasuk deteksi hoaks, analisis sentimen, dan klasifikasi teks. Sebagai contoh, IndoBERT mencapai akurasi 90% dalam mendeteksi berita hoaks [4]. Selain itu, model ini sering digunakan untuk mengatasi disinformasi di media sosial dan analisis pola teks berbahasa Indonesia [5], [6].

IndoBERT tersedia dalam beberapa varian, seperti *IndoBERT-Large*, yang lebih besar, dan *IndoBERT-Lite*, yang lebih kecil dan lebih efisien. Model ini memberikan solusi yang fleksibel untuk berbagai aplikasi NLP yang berfokus pada bahasa Indonesia.

C. Cendol

Cendol adalah model yang dirancang untuk menangani kompleksitas bahasa Indonesia, termasuk idiom, ungkapan lokal, dan variasi dialek yang sering muncul dalam komunikasi sehari-hari maupun di media sosial. Dengan memanfaatkan arsitektur *deep learning*, Cendol mampu memahami konteks dan makna teks dengan lebih akurat. Model ini telah terbukti efektif dalam tugas-tugas seperti deteksi ujaran kebencian dan klasifikasi teks. Selain itu, Cendol mendukung analisis multimodal, yang memungkinkan penggabungan analisis teks dan gambar, misalnya dalam mendeteksi makna meme.

Cendol dibangun berdasarkan dua arsitektur utama, yaitu mT5 dan LLaMA-2, dengan varian sebagai berikut:

- *mT5-based Cendol*
  - Cendol-Mini: 300 juta parameter (*mT5-small*)
  - Cendol-Base: 580 juta parameter (*mT5-base*)
  - Cendol-Large: 1,2 miliar parameter (*mT5-large*)
  - Cendol-XL: 3,7 miliar parameter (*mT5-XL*)
  - Cendol-XXL: 13 miliar parameter (*mT5-XXL*)
- *LLaMA-2-based Cendol*
  - Cendol-LLaMA2-7B: 7 miliar parameter
  - Cendol-LLaMA2-13B: 13 miliar parameter

Cendol memiliki dua varian utama: *Cendol-Instruct* dan *Cendol-Chat*. *Cendol-Instruct* dirancang untuk tugas-tugas yang terstruktur, seperti analisis sentimen dan terjemahan mesin, sementara *Cendol-Chat* dioptimalkan untuk percakapan interaktif dan berbasis konteks kehidupan sehari-hari.

Model ini dirancang untuk menangkap nuansa bahasa Indonesia dengan lebih baik dan dapat diterapkan dalam berbagai tugas NLP, seperti klasifikasi teks dan analisis sentimen. Dalam penelitian ini, salah satu varian yang digunakan adalah *Cendol-mt5-small-inst*, yang dioptimalkan untuk tugas-tugas spesifik. Model ini memberikan solusi untuk menangani tantangan bahasa lokal yang tidak dapat ditangani secara optimal oleh model umum [7].

III. METODE

Fine-tuning dan *hyperparameter tuning* adalah dua teknik penting dalam optimasi model *machine learning* yang dirancang untuk meningkatkan performa model.

A. *Fine-tuning*

Berbeda dengan pendekatan *feature-based* yang tidak memodifikasi model yang telah dilatih, *fine-tuning* melibatkan pelatihan lanjutan pada model *pretrained* dengan menggunakan dataset yang lebih kecil dan spesifik untuk penyesuaian terhadap kebutuhan tertentu [8]. Tujuannya adalah untuk mempertahankan kemampuan awal model sambil mengoptimalkannya untuk tugas-tugas tertentu. Pendekatan ini menjadi solusi ideal dalam situasi dengan keterbatasan sumber daya komputasi atau data relevan yang terbatas.

Tabel I: *Hyperparameter* yang Digunakan pada Model NLP.

Hyperparameter	Nilai
Epoch	5,10,15,20
Sequence Length	256
Optimizers	Adam
Batch Size	2,4,8
Learning Rate	0.1 ; 0.01 ; 0.001
Metrics	*Categorical Accuracy*
Losses	*Categorical Crossentropy*

Hyperparameter adalah parameter yang digunakan untuk mengatur konfigurasi model sebelum proses pelatihan dimulai. Nilai-nilai ini sangat memengaruhi pola pembelajaran model serta kemampuan generalisasinya terhadap data baru.

B. *Evaluasi Model*

Confusion matrix adalah alat visualisasi yang merepresentasikan hubungan antara hasil prediksi model dan data aktual. Matriks ini terdiri dari empat komponen utama: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN), yang digunakan untuk menilai kinerja model [9]. Gambar 1 menunjukkan contoh representasi *confusion matrix* pada kasus klasifikasi biner.

Berdasarkan *confusion matrix*, berbagai metrik evaluasi kinerja model dapat dihitung, antara lain:

- ***Accuracy*:** Mengukur persentase prediksi yang benar terhadap total prediksi. Persamaan *accuracy* 1

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

di mana TP dan TN menunjukkan jumlah prediksi yang benar, sedangkan FP dan FN menunjukkan jumlah prediksi yang salah.

	Total Pop	Actual Class: Positive Class	Actual Class: Negative Class
Predicted Class	Predicted Positive Class	True Positive	False Positive
Predicted Class	Predicted Negative Class	False Negative	True Negative

GAMBAR 1:
Ilustrasi *confusion matrix*.

- ***Precision*:** Menunjukkan sejauh mana model akurat dalam mengidentifikasi *hate comments*. Persamaan *precision* 2

$$\text{P} = \frac{TP}{TP + FP} \quad (2)$$

dengan TP adalah jumlah *true positive*, dan FP adalah jumlah *false positive*.

- ***Recall*:** Menilai kemampuan model dalam mendeteksi semua *hate comments*. Persamaan *recall* 3

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

dengan TP adalah jumlah *true positive*, dan FN adalah jumlah *false negative*.

- ***F1-score*:** Merupakan rata-rata harmonik antara *precision* dan *recall*. Persamaan *f1-score* 4

$$\text{F1-score} = \frac{2 \times \text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad (4)$$

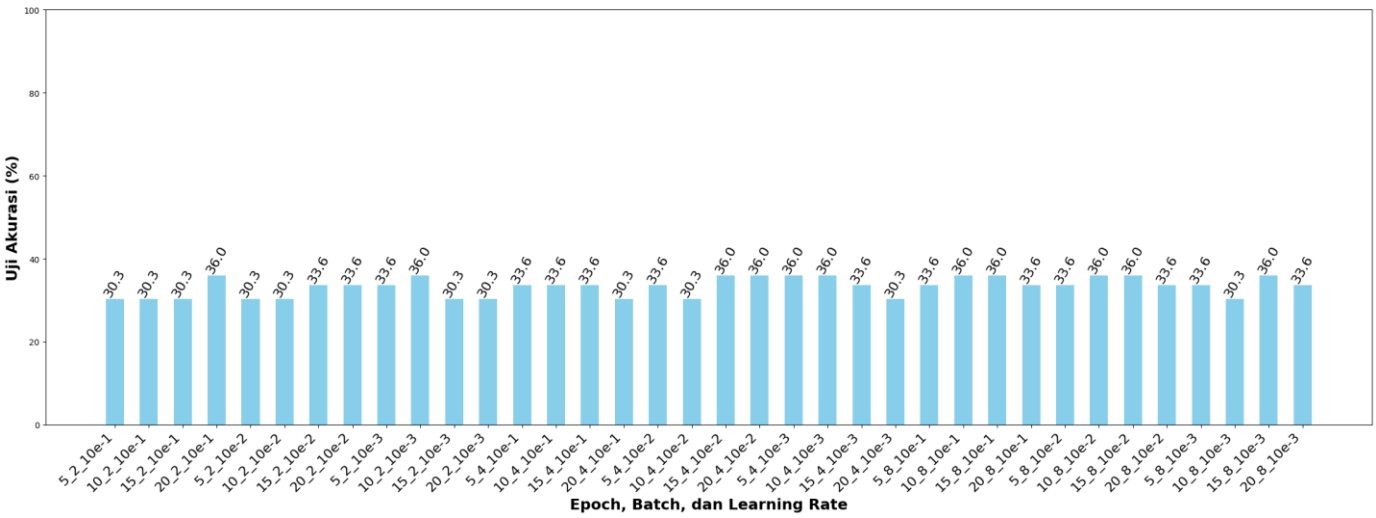
F1-score memberikan evaluasi yang lebih menyeluruh dengan mempertimbangkan keseimbangan antara *precision* dan *recall*.

IV. HASIL DAN PEMBAHASAN

A. IndoBERT

Berikut merupakan hasil *test accuracy* pada model IndoBERT. Hasil akurasi pengujian model IndoBERT yang dianalisis berdasarkan kombinasi parameter *epoch*, *batch size*, dan *learning rate*. Nilai akurasi ditampilkan dalam bentuk diagram batang, yang mengilustrasikan variasi hasil model pada setiap kombinasi parameter tersebut.

Gambar 2 menunjukkan hasil uji akurasi model IndoBERT dengan variasi *epoch*, *batch size*, dan *learning rate*. Pada sumbu vertikal, akurasi uji (%) diukur, sedangkan sumbu horizontal menampilkan kombinasi parameter tersebut. Model diuji dengan beberapa konfigurasi yang meliputi *epoch* 5, 10, 15, 20, *batch size*, dan *learning rate* 10^{-1} - 10^{-3}.

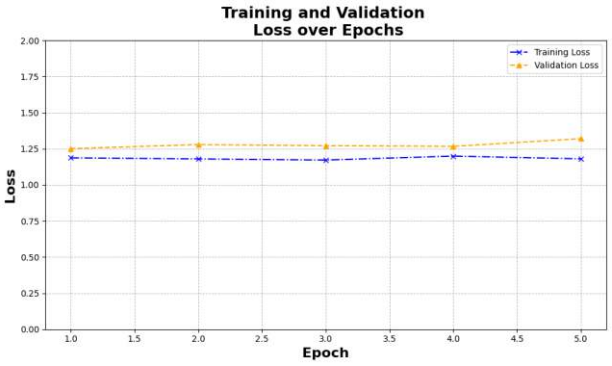


GAMBAR 2:
Test accuracy pada model IndoBERT.

Dari grafik ini, menunjukkan variasi akurasi antara 30,3% hingga 36,0%, dengan kombinasi tertentu memberikan performa lebih baik. Akurasi tertinggi sebesar 36,0% diperoleh pada konfigurasi *epoch* 20, *batch size* 5, dan l*earning rate* 10^{-2}. Namun, sebagian besar konfigurasi menghasilkan akurasi sekitar 30%-33,6%. Variasi *learning rate* dan *epoch* mempengaruhi peningkatan performa model, sedangkan *batch size* memiliki pengaruh yang lebih konsisten terhadap hasil uji.

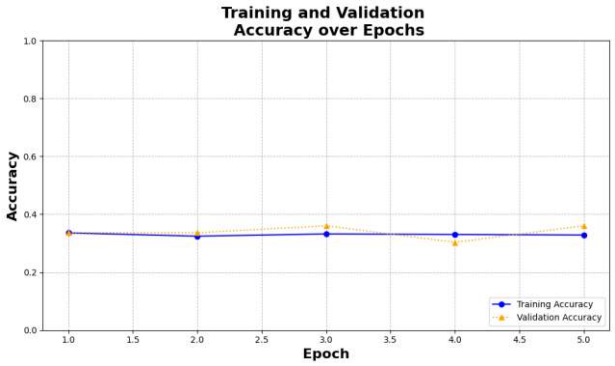
Gambar 3 menunjukkan menunjukkan grafik *loss* untuk data pelatihan dan validasi selama 5 *epoch* dalam proses pelatihan model. Pada sumbu horizontal, grafik menampilkan jumlah *epoch* dari 1 hingga 5, sedangkan sumbu vertikal menunjukkan nilai loss dengan rentang dari 0 hingga 2. Garis biru putus-putus dengan tanda bintang menggambarkan *training loss*, sementara garis oranye putus-putus dengan tanda segitiga menunjukkan *validation loss*. Selama proses pelatihan, nilai *training loss* terlihat relatif stabil di sekitar angka 1.2, dengan sedikit fluktuasi namun cenderung konstan. Di sisi lain, *validation loss* memiliki *tren* yang sedikit meningkat, dimulai dari sekitar 1.25 pada *epoch* 1 dan naik menjadi 1.33 pada *epoch* ke-5. Perbedaan antara kedua *loss* ini menunjukkan adanya *gap* yang kecil, yang mengindikasikan performa model pada data pelatihan dan validasi relatif konsisten meskipun ada kecenderungan *overfitting* kecil di akhir pelatihan. Secara keseluruhan, grafik ini menunjukkan bahwa model dapat belajar dengan stabil, tetapi mungkin perlu dilakukan penyesuaian hyperparameter atau regularisasi tambahan untuk mengurangi peningkatan *validation loss* yang terlihat di *epoch* akhir.

Sedangkan Gambar 4 menunjukkan perkembangan akurasi pelatihan dan akurasi validasi selama 5 *epoch*. Pada sumbu horizontal (x) ditampilkan jumlah *epoch*, sedangkan sumbu vertikal (y) merepresentasikan nilai akurasi dalam rentang 0 hingga 1. Selama 5 *epoch*, akurasi *training accuracy* dan *validation accuracy* menunjukkan pola yang relatif stabil tanpa peningkatan yang signifikan. Pada awal *epoch*, akurasi pelatihan dan validasi berada di sekitar 0,4. Meskipun, terdapat sedikit fluktuasi pada beberapa titik, *tren* keseluruhan menunjukkan bahwa model tidak mengalami perbaikan substansial dari *epoch* pertama hingga kelima. Sehingga, grafik mengindikasikan bahwa model mungkin mengalami stagnasi, yang dapat disebabkan oleh kurangnya pembelajaran lebih lanjut, seperti parameter model yang tidak optimal, arsitektur yang kurang kompleks, atau kebutuhan untuk menyesuaikan *hyperparameter*. Hal ini terlihat dari garis akurasi validasi yang hampir selalu berdekatan dengan garis akurasi pelatihan, yang berarti tidak ada tanda-tanda *overfitting* yang jelas, tetapi juga tidak ada perbaikan performa yang nyata.



GAMBAR 3:
Loss pada model IndoBERT dengan *epoch* 5, *learning rate* 10e-3, dan *batch size* 4.

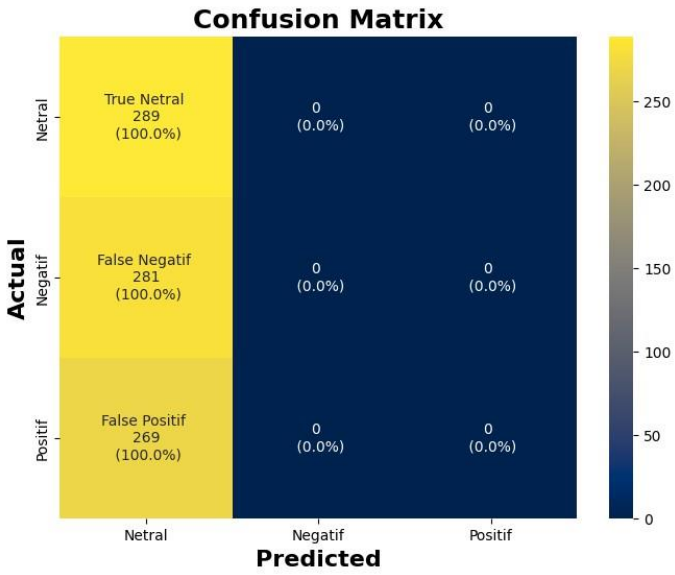
Gambar 5 menunjukkan *confusion matrix* pada model IndoBERT dengan tiga label, yaitu positif, negatif, dan netral, yang menggambarkan performa model dalam mengklasifikasikan data ke dalam ketiga label tersebut. Matriks ini menggambarkan hasil prediksi model terhadap data aktual. Setiap sel dalam matriks menunjukkan jumlah prediksi yang dibuat oleh model untuk setiap kategori. Pada baris pertama,



GAMBAR 4:
Accuracy pada model IndoBERT dengan *epoch* 5, *learning rate* 10e-3, dan *batch size* 4.

model berhasil mengklasifikasikan semua data aktual kategori netral dengan benar sebagai netral, dengan jumlah 289 data (100%). Namun, tidak ada prediksi model untuk kategori negatif atau positif pada data yang sebenarnya netral, sehingga sel lainnya di baris ini bernilai nol. Baris kedua menunjukkan bahwa model salah memprediksi semua data aktual kategori negatif. Sebanyak 281 data negatif sepenuhnya salah diklasifikasikan, dengan rincian bahwa semuanya dianggap sebagai netral oleh model. Selain itu, tidak ada prediksi kategori negatif maupun positif pada data aktual kategori ini. Baris ketiga mencerminkan hasil prediksi untuk data aktual kategori positif. Sebanyak 269 data positif juga sepenuhnya salah diklasifikasikan oleh model. Sehingga semuanya diprediksi sebagai netral, sama seperti sebelumnya, tidak ada prediksi yang benar untuk kategori ini. Secara keseluruhan, *confusion matrix* ini mengindikasikan bahwa model hanya dapat memprediksi kategori netral dengan benar, tetapi gagal sepenuhnya dalam mengenali data kategori negatif dan positif. Hal ini menunjukkan bahwa model memiliki bias kuat terhadap kategori netral dan memerlukan evaluasi serta perbaikan signifikan untuk meningkatkan performa pada kategori lainnya.

Gambar 6 menunjukkan laporan klasifikasi pada model IndoBERT dengan tiga label yaitu netral, negatif, dan positif. Laporan ini menyajikan metrik evaluasi seperti *accuracy*, *precision*, *recall*, *F1-Score*, dan *support* untuk setiap label. Pada kelas netral, model mencapai *precision* sebesar 0,34, *recall* sebesar 1,00, dan *F1-Score* sebesar 0,51 dengan total data sebanyak 289. Selain itu, untuk kelas negatif, model menunjukkan *precision*, *recall*, dan *F1-Score* sebesar 0,00 dengan total data sebanyak 281. Kelas positif juga memiliki *precision*, *recall*, dan *F1-Score* sebesar 0,00 dari total 269 data. Secara keseluruhan, akurasi model tercatat sebesar 0,34 dari 839 data. Nilai rata-rata secara makro (*macro avg*) menunjukkan *precision* sebesar 0,11, *recall* sebesar 0,33, dan *F1-Score* sebesar 0,17. Sementara itu, rata-rata berbobot (*weighted avg*) menghasilkan *precision* sebesar 0,12, *recall* sebesar 0,34, dan *F1-Score* sebesar 0,18. Laporan ini menunjukkan bahwa model bekerja lebih baik pada kelas netral, namun kinerjanya sangat buruk pada kelas negatif dan positif. Model memerlukan evaluasi dan perbaikan agar dapat menangani data secara lebih seimbang.



GAMBAR 5:
Confusion Matrix pada model IndoBERT.

Laporan Klasifikasi:

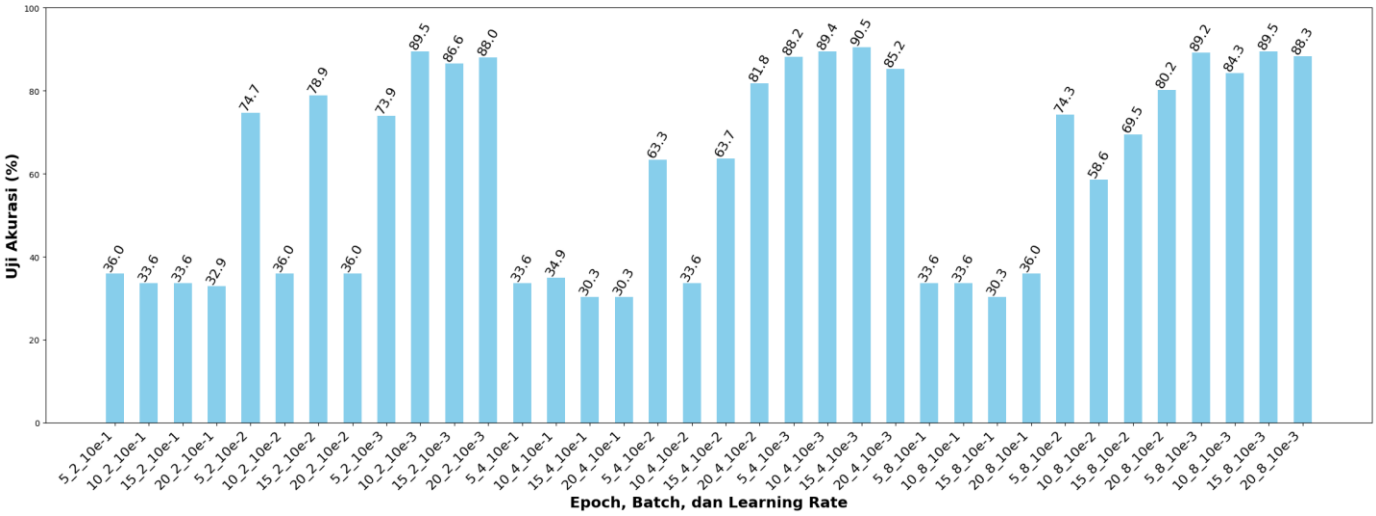
	precision	recall	f1-score	support
Netral	0.34	1.00	0.51	289
Negatif	0.00	0.00	0.00	281
Positif	0.00	0.00	0.00	269
accuracy			0.34	839
macro avg	0.11	0.33	0.17	839
weighted avg	0.12	0.34	0.18	839

GAMBAR 6:
Laporan klasifikasi pada model IndoBERT.

B. *Cendol*

Berikut merupakan hasil *test accuracy* pada model Cendol. Hasil akurasi pengujian model Cendol yang dianalisis berdasarkan kombinasi parameter *epoch*, *batch size*, dan *learning rate*. Nilai akurasi ditampilkan dalam bentuk diagram batang, yang mengilustrasikan variasi hasil model pada setiap kombinasi parameter tersebut.

Gambar 7 menunjukkan grafik akurasi pengujian model Cendol berdasarkan variasi kombinasi parameter *epoch*, *batch size*, dan *learning rate*. Sumbu vertikal menunjukkan akurasi pengujian dalam bentuk persentase (%), sementara sumbu horizontal merepresentasikan berbagai kombinasi parameter yang digunakan selama pelatihan. Hasil menunjukkan bahwa akurasi model tercatat dalam persentase dan bervariasi tergan- tung pada parameter yang digunakan. Pada kombinasi pertama *epoch* 5, *batch size* 10, dan *learning rate* 10^{-1}, akurasi hanya mencapai 36%. Akurasi tertinggi sebesar 90,4% dicapai saat parameter diatur menjadi 15 *epoch*, *batch size* 10, dan learning rate 0,3. Beberapa kombinasi parameter lain seperti 15 *epoch*, *batch size* 20, dan *learning rate* 10^{-3} juga memberikan akurasi

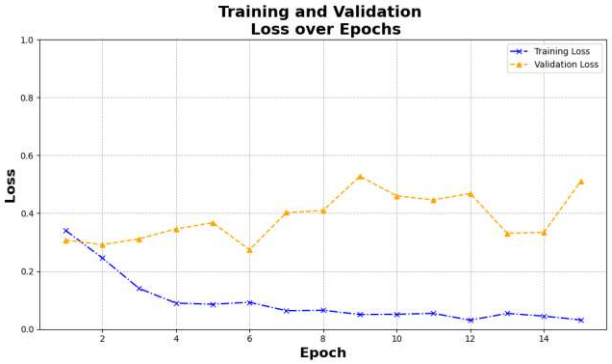


GAMBAR 7:
Test accuracy pada model Cendol.

yang mendekati maksimal sebesar 89,5%. Selain itu, beberapa kombinasi seperti *batch size* 5 dengan *learning rate* 10^{-2} menghasilkan akurasi yang lebih rendah. Grafik ini memperlihatkan bahwa pemilihan *epoch*, *batch size*, dan *learning rate* memengaruhi akurasi model secara signifikan.

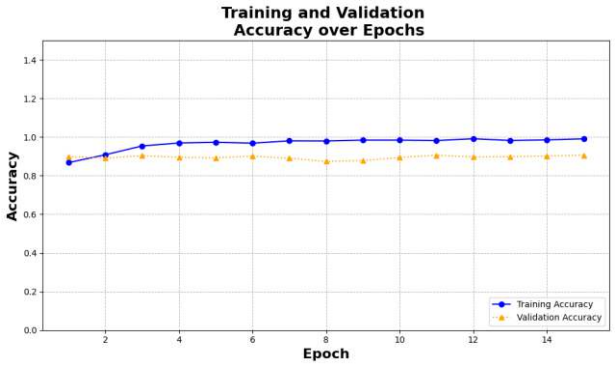
Gambar 8 menunjukkan grafik *training loss* dan *validation loss* terhadap *epoch*. Sumbu horizontal menunjukkan jumlah *epoch* yang digunakan selama proses pelatihan, sedangkan sumbu vertikal merepresentasikan nilai *loss*. Pada grafik, garis biru dengan penanda ”X” menunjukkan *training loss* yang konsisten menurun seiring bertambahnya *epoch*. Nilai *training loss* dimulai dari 0.35 pada *epoch* pertama dan terus turun hingga mendekati 0.05 pada *epoch* terakhir, menunjukkan bahwa model berhasil belajar dari data pelatihan dengan baik. Garis oranye dengan penanda segitiga menunjukkan *validation loss*. Nilai *validation loss* tampak fluktuatif sepanjang *epoch*. Pada awalnya, nilai *validation loss* mendekati 0.3, kemudian mengalami kenaikan dan penurunan yang tidak stabil, dengan puncak mencapai sekitar 0.55 pada *epoch* 8 dan 15. Hal ini mengindikasikan bahwa model mungkin mengalami *overfitting* terhadap data pelatihan, karena *validation loss* tidak mengikuti pola penurunan seperti *training loss*. Secara keseluruhan, grafik ini menggambarkan kinerja model selama proses pelatihan dan validasi, dengan pola stabil pada *training loss* namun adanya ketidakseimbangan pada *validation loss*. Model perlu dievaluasi lebih lanjut untuk menangani potensi *overfitting* yang terjadi.

Sedangkan Gambar 9 menampilkan grafik *training accuracy* dan *validation accuracy* terhadap epoch. Sumbu horizontal merepresentasikan jumlah epoch, sedangkan sumbu vertikal menunjukkan tingkat akurasi model. Garis biru dengan penanda lingkaran menggambarkan *training accuracy*. Nilai *training accuracy* mengalami peningkatan dari sekitar 0.87 pada *epoch* pertama hingga mendekati 1.0 pada *epoch* terakhir. *Tren* ini menunjukkan bahwa model berhasil belajar dengan baik dari data pelatihan. Garis oranye dengan penanda segitiga menunjukkan Validation Accuracy. Nilai *validation accuracy* memulai dari sekitar 0.89 pada *epoch* pertama dan mengalami sedikit fluktuasi sepanjang proses pelatihan, tetap berada di kisaran 0.88 hingga 0.90. Grafik ini menunjukkan bahwa akurasi validasi cenderung stabil, namun tidak meningkat signifikan seperti *training accuracy*. Perbedaan kecil antara *training accuracy* dan *validation accuracy* mengindikasikan potensi *overfitting*. Model belajar dengan sangat baik pada data pelatihan tetapi kurang generalisasi terhadap data validasi. Evaluasi lanjutan mungkin diperlukan untuk mengatasi masalah tersebut.



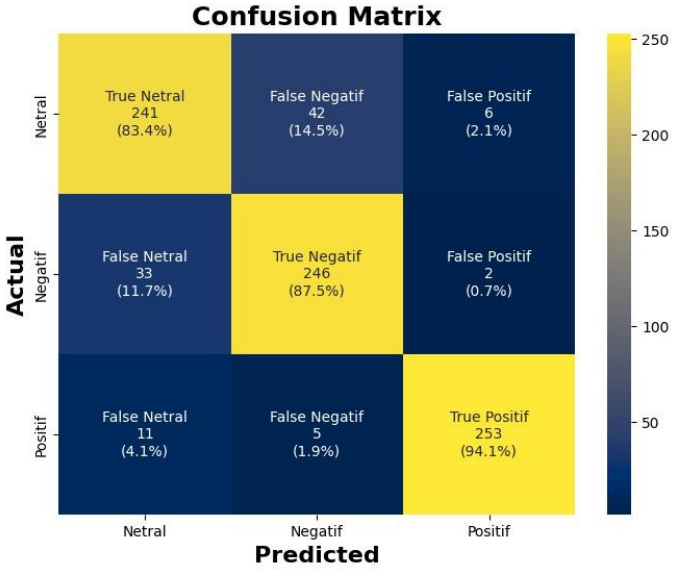
GAMBAR 8:
Loss pada model Cendol dengan *epoch* 15, *learning rate* 10^{-3}, dan *batch size* 4.

Gambar 10 menunjukkan *confusion matrix* dari hasil klasifikasi model yang mengelompokkan data ke dalam tiga kategori yaitu netral, negatif, dan positif. Pada kelas netral, model memprediksi dengan benar sebanyak 241 sampel (83,4%). Namun, terdapat 42 sampel (14,5%) yang diprediksi sebagai negatif dan 6 sampel (2,1%) yang diprediksi sebagai positif. Pada kelas negatif, model berhasil mengklasifikasikan den-



GAMBAR 9:
Accuracy pada model Cendol dengan *epoch* 15, *learning rate* 10^{-3}, dan *batch size* 4.

gan benar 246 sampel (87,5%). Namun, 33 sampel (11,7%) salah diprediksi sebagai netral, sedangkan 2 sampel (0,7%) diprediksi menjadi positif. Sedangkan pada kelas positif, model memiliki akurasi yang tinggi dengan 253 sampel (94,1%) diprediksi dengan benar. Namun, terdapat kesalahan prediksi pada 11 sampel (4,1%) yang diklasifikasikan sebagai Netral dan 5 sampel (1,9%) yang diprediksi sebagai negatif. Secara keseluruhan, model memiliki performa yang cukup baik dengan tingkat kesalahan prediksi yang kecil pada semua kelas. Sehingga angka yang ditunjukkan mencakup jumlah sampel beserta persentase dari setiap kategori prediksi.



GAMBAR 10:
Confusion Matrix pada model Cendol.

Gambar 11 menunjukkan laporan klasifikasi pada model Cendol dengan tiga label yaitu netral, negatif, dan positif. Pada kelas netral, model memiliki *precision* sebesar 0.85, *recall* sebesar 0.83, dan *F1-Score* sebesar 0.84 dengan total 289 sampel. Pada kelas negatif menunjukkan *precision* sebesar 0.84, *recall* sebesar 0.88, dan *F1-Score* sebesar 0.86 dari 281 sampel. Sedangkan pada kelas positif, model mencapai performa tertinggi dengan *precision* sebesar 0.97, *recall* sebesar 0.94, dan *F1-Score* sebesar 0.95 dari 269 sampel. Secara keseluruhan, model memperoleh akurasi sebesar 0.88 untuk 839 sampel. Rata-rata makro (*macro avg*) dan rata-rata berbobot (*weighted avg*) untuk *precision*, *recall*, dan *F1-Score* masing-masing mencapai 0.88. Laporan ini menunjukkan performa model yang konsisten dengan akurasi tinggi, terutama pada kelas positif yang memiliki nilai evaluasi terbaik. Sedangkan kelas netral dan negatif juga memiliki hasil yang baik meskipun sedikit di bawah kelas positif.

Laporan Klasifikasi:

	precision	recall	f1-score	support
Netral	0.85	0.83	0.84	289
Negatif	0.84	0.88	0.86	281
Positif	0.97	0.94	0.95	269
accuracy			0.88	839
macro avg	0.88	0.88	0.88	839
weighted avg	0.88	0.88	0.88	839

Gambar 11: Laporan klasifikasi pada model Cendol pada *epoch* 15, *batch size* 4, dan *learning rate* 10^{-1}.

V. KESIMPULAN

Penelitian ini menganalisis performa model IndoBERT dan Cendol dalam mendeteksi komentar kebencian yang berkaitan dengan *cyberbullying*. Dataset yang digunakan terdiri dari 3875 data, yang diproses melalui tahap pembersihan dan pelabelan manual. Kedua model dievaluasi menggunakan berbagai konfigurasi *hyperparameter*, seperti *epoch*, *learning rate*, *batch size*, serta diukur performansinya berdasarkan akurasi, presisi, *recall*, dan *F1-Score*.

Berdasarkan hasil eksperimen, model Cendol mencapai akurasi tertinggi sebesar 90,5% pada konfigurasi optimal *epoch* 15, *batch size* 4, dan *learning rate* 10^{-3}. Hasil klasifikasi Cendol menunjukkan nilai *precision* 0.85 untuk kategori Netral, 0.84 untuk negatif, dan 0.97 untuk Positif, dengan nilai *recall* masing-masing 0.83, 0.88, dan 0.94. F1-Score yang dihasilkan untuk kategori netral, negatif, dan positif adalah 0.84, 0.86, dan 0.95. Model Cendol menunjukkan performa yang baik dengan akurasi keseluruhan 88% dan nilai rata-rata makro dan tertimbang masing-masing 0.88.

Sementara itu, model IndoBERT menunjukkan akurasi hanya sebesar 36% pada konfigurasi terbaik *epoch* 5, *batch size* 4, dan *learning rate* 10^{-3}. Laporan klasifikasi IndoBERT menunjukkan hasil yang sangat kurang memadai, dengan nilai *precision* 0.34 untuk kategori netral, dan 0.00 untuk kategori negatif dan positif. Nilai *recall* untuk kategori netral adalah 1.00, tetapi untuk kategori negatif dan positif adalah 0.00. *F1-Score* untuk kategori Netral adalah 0.51, namun untuk kategori negatif dan positif tidak terhitung. Akurasi keseluruhan model IndoBERT hanya mencapai 34%, dengan nilai rata-rata makro dan tertimbang yang sangat rendah, masing-masing 0.17 dan 0.18.

Hasil ini menunjukkan bahwa model Cendol lebih efektif dalam mendeteksi komentar kebencian, khususnya pada konteks bahasa informal dan lokal. Sebaliknya, performa IndoBERT dalam eksperimen ini terbatas, yang mungkin disebabkan oleh ketidaksesuaian antara model dengan dataset yang digunakan.

Penelitian ini menggarisbawahi pentingnya optimasi *hyperparameter* dalam meningkatkan kemampuan deteksi model, sekaligus menyediakan solusi berbasis teknologi untuk mengurangi dampak negatif *cyberbullying* di berbagai *platform* media sosial, seperti X, Facebook, Instagram, dan TikTok. Namun, adanya potensi *overfitting* pada model tertentu menjadi perhatian untuk penelitian di masa mendatang.

REFERENSI

[1] Y. A. Cahyadi, "Kampanye pemilu 2024, ujaran kebencian terhadap kelompok minoritas meningkat," *AJI Indonesia*, Feb. 2024, Accessed: 2024-10-18. [Online]. Available: https://aji.or.id/informasi/kampanye- pemilu-2024- ujaran- kebencian- terhadap- kelompok- minoritas-meningkat.

[2] N. Muhamad, "Twitter, medsos dengan ujaran kebencian terbanyak pada kampanye pemilu 2024," *Databoks Premium Lite*, Feb. 2024, Accessed: 2024-10-18. [Online]. Available: https : / / databoks . katadata . co . id / teknologi - telekomunikasi / statistik / 16c6c45ef50c346 / twitter - medsos - dengan - ujaran - kebencian - terbanyak - pada-kampanye-pemilu-2024.

[3] P. Nabila and E. B. Setiawan, "Adam and adamw optimization algorithm application on bert model for hate speech detection on twitter," *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 346–351, 2024. [Online]. Available: https : / / api . semanticscholar.org/CorpusID:272432470.

[4] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax news detection analysis using indobert deep learning methodology," *2022 10th International Conference on Information and Communication Technology (ICoICT)*, pp. 368–373, 2022. [Online]. Available: https : / / api . semanticscholar.org/CorpusID:253047025.

[5] A. B. Y. A. Putra and Y. Sibaroni, "Disinformation detection on 2024 indonesia presidential election using indobert," *2023 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 350–355, 2023. [Online]. Available: https : / / api . semanticscholar . org / CorpusID:264293048.

[6] L. R. Aini, E. Nurfadhilah, A. Jarin, A. Santosa, and M. T. Uliniansyah, "Enhancing sentiment analysis models through multi-technique data augmentation: A study with indobert," *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 137–142, 2023. [Online]. Available: https: //api.semanticscholar.org/CorpusID:264294774.

[7] H. Face, *Cendol: Open instruction-tuned generative large language models for indonesian languages*, Accessed: 2024-12-09, 2024. [Online]. Available: https:// huggingface.co/indonlp/cendol.

[8] L. Craig, *What is fine-tuning in machine learning and ai?* Accessed: 10-11-2024, Jul. 2024. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/ fine-tuning.

[9] A. R. Hanum, I. A. Zetha, S. C. Putri, *et al.*, "Analisis kinerja algoritma klasifikasi teks bert dalam mendeteksi berita hoaks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, pp. 537–546, 2024.