UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

IndoBERTweet – BILSTM untuk Deteksi Cyberbullying Pada Teks Media Sosial Campuran Bahasa Indonesia - Inggris

Oleh :

Mochammad Wahyu Alvy Kusuma
(221080200117)

Dosen Pembimbing :
Dr. Mochamad Alfan Rosid, S.Kom., M.Kom.

Dosen Penguji 1:
Sukma Aji, S.T., M.Kom

Dosen Penguji 2:
Suhendro Busono, S.ST., M.Kom

Program Studi Informatika
Universitas Muhammadiyah Sidoarjo

2025

Latar Belakang

Laporan We are Social menunjukkan, jumlah pengguna aktif media sosial di Indonesia sebanyak 167 juta orang pada januari 2023,dengan platform X menjadi salah satunya paling banyak digunakan.

Perkembangan media sosial telah menjadi bagian penting untuk kehidupan sehari hari, memungkinkan interaksi dan berbagi pendapat secara real-time

Namun, kemudahan akses yang ditawarkan juga membawa dampak negatif, salah satunya meningkatnya kasus Cyberbullying atau perudungan siber.

Dampak negatif yang ditimbulkan dari Cyberbullying sangatah serius mulai dari gangguan psikologi, depersi, hingga kasus bunuh diri

Model IndoBERTweet menawarkan solusi efektif untuk secara otomatis mengidentifikasi dan menyaring komentar maupun ciutan yang mengandung Cyberbullying

Rumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah:
“Bagaimana mengembangkan dan mengevaluasi model IndoBERTweet dalam mendeteksi cyberbullying pada komentar di platform X”.

Tujuan

Untuk mengembangkan dan mengevaluasi model IndoBERTweet untuk mendeteksi cyberbullying pada komentar di platform X.

Penelitian Terdahulu

P. Hanif Zakaria, D. Nurjannah and H. Nurrahmi (2023)

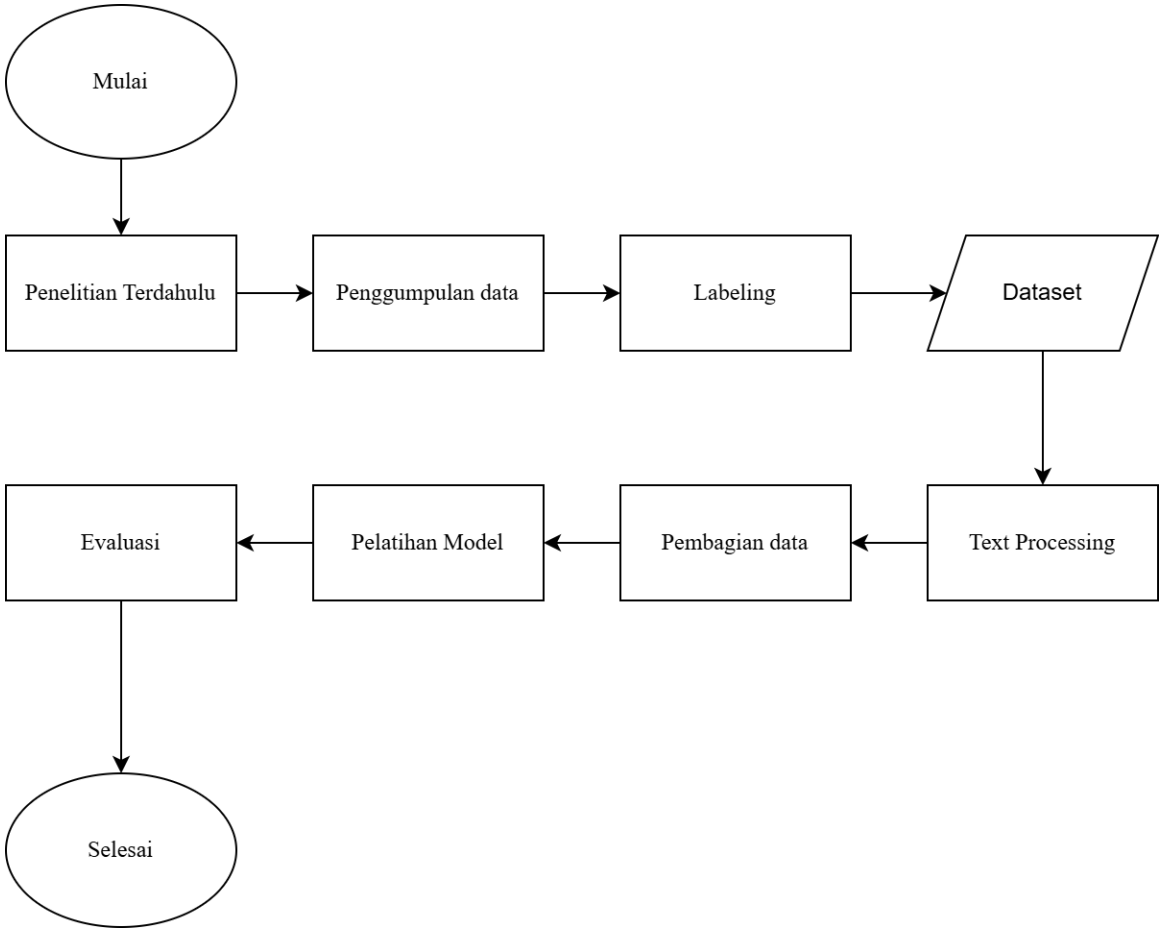
Judul : Misogyny Text Detection on Tiktok Social Media in Indonesian Using the Pre-trained Language Model IndoBERTweet

J. Forry , O. Andry Chowanda (2023)

Judul : Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter

Kemsimpulan : Penelitian sebelumnya jarang membahas bahasa campuran Indonesia–Inggris. Solusi yang ditawarkan yaitu menggunakan dataset campuran dan model IndoBERTweet yang dioptimalkan untuk bahasa Indonesia.

Tahapan Penelitian



Rancangan Penelitian

- Pengumpulan Dataset

Tweet Comment
Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @ DivHumas_ Polri ini ditindaklanjuti . https://t.co/5GtpY1FHCm
Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa

Data penelitian ini diambil dari komentar pada platform X dengan menggunakan API.

Rancangan Penelitian

- Pelabelan Dataset

Tweet Comment	Label
Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @ DivHumas_ Polri ini ditindaklanjuti . https://t.co/5GtpY1FHCm	1
Gibran Datang ke Rumahnya Rocky Gerung: Saya Kasih Kopi Oke You Bicara Anak Muda! https://t.co/ewhKdZIGGa	0

1 (bullying)

0 (nonbullying)

Tahap selanjutnya pelabelan data secara manual.

Rancangan Penelitian

- Text Processing

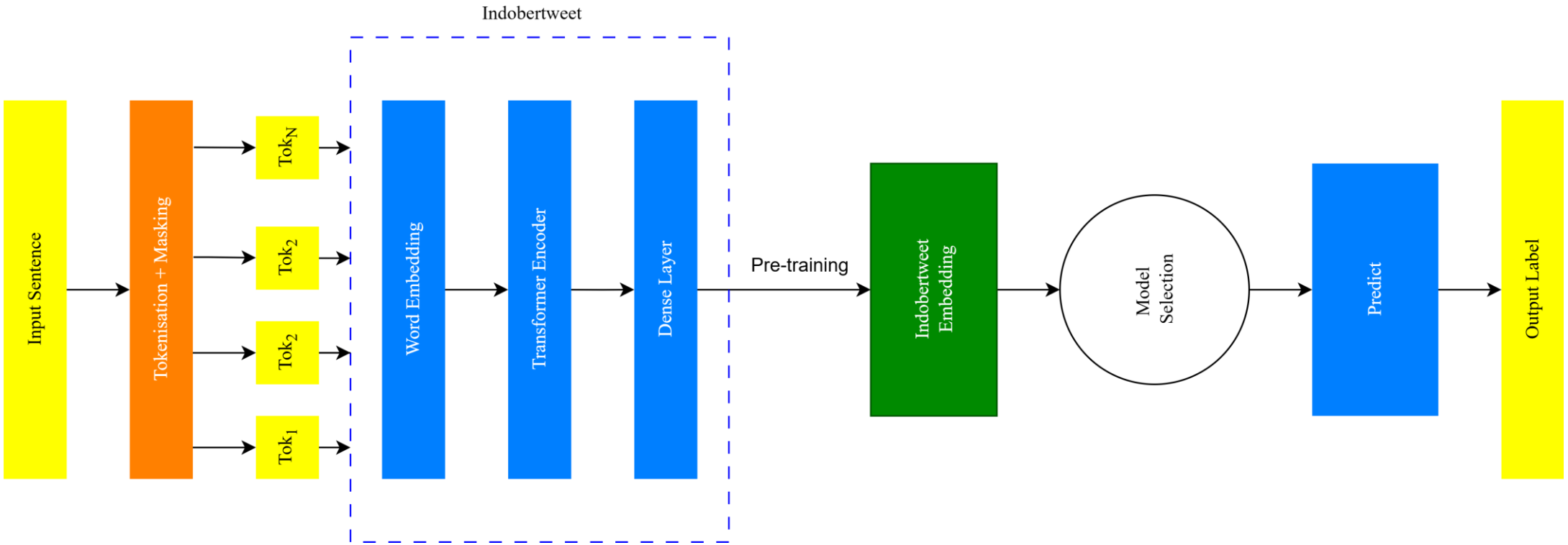
Tweet Comment	Sesudah
Goblok ini Hoaxs ..Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs .. Tolong @ DivHumas_ Polri ini ditindaklanjuti . https://t.co/5GtpY1FHCm	Goblok ini Hoaxs Tolong Kalo Jadi Pendukung Anies terus Jangan Gilak Bikan Hoaxs Tolong ini ditindaklanjuti

Kemudian tahap text processing.

- Pembagian Data

Data dibagi menjadi 3 yaitu data latih sebesar 80 % (8.752), data uji 10% (2.736), dan data validasi 10% (2.189)

Arsitektur Model



Tahap Pengujian

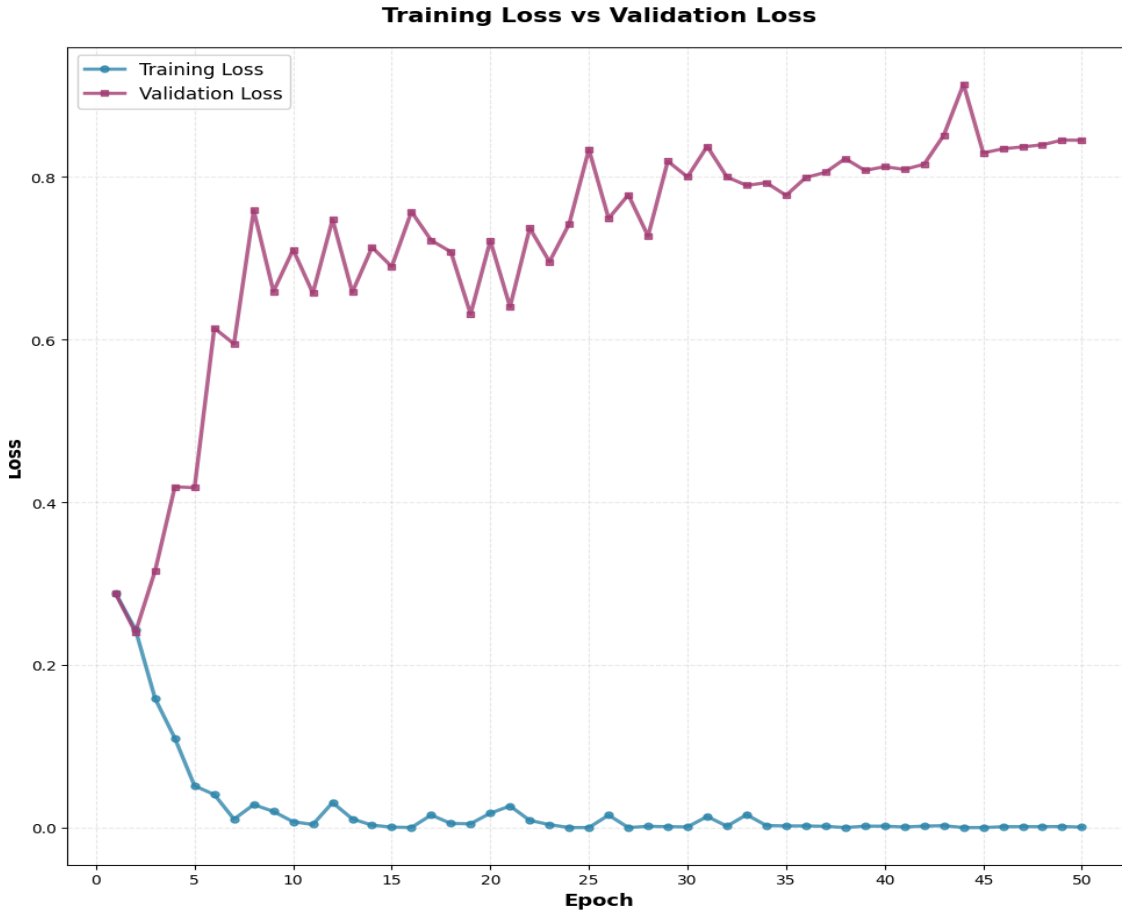
Pengujian ini menerapkan 3 scenario pengujian dengan variasi learning rate dan batch size. Skenario 1 sebagai baseline menggunakan learning rate 2e-5 dan batch size 16. Skenario 2 meningkatkan learning rate menjadi 3e-5 dengan batch size tetap 16. Skenario 3 menggunakan batch size 32 dengan learning rate kembali ke 2e-5. semua scenario dijalankan selama 50 epoch dengan data latih dan validasi yang sama.

Evaluasi Model

Pada tahap evaluasi model, model yang telah dikembangkan diuji untuk menilai kinerjanya. Proses evaluasi dilakukan dengan menggunakan confusion matrix, serta analisis kinerja model diukur melalui metrik evaluasi yaitu accuracy, precision, recall, dan f1-score.

Hasil & Pembahasan

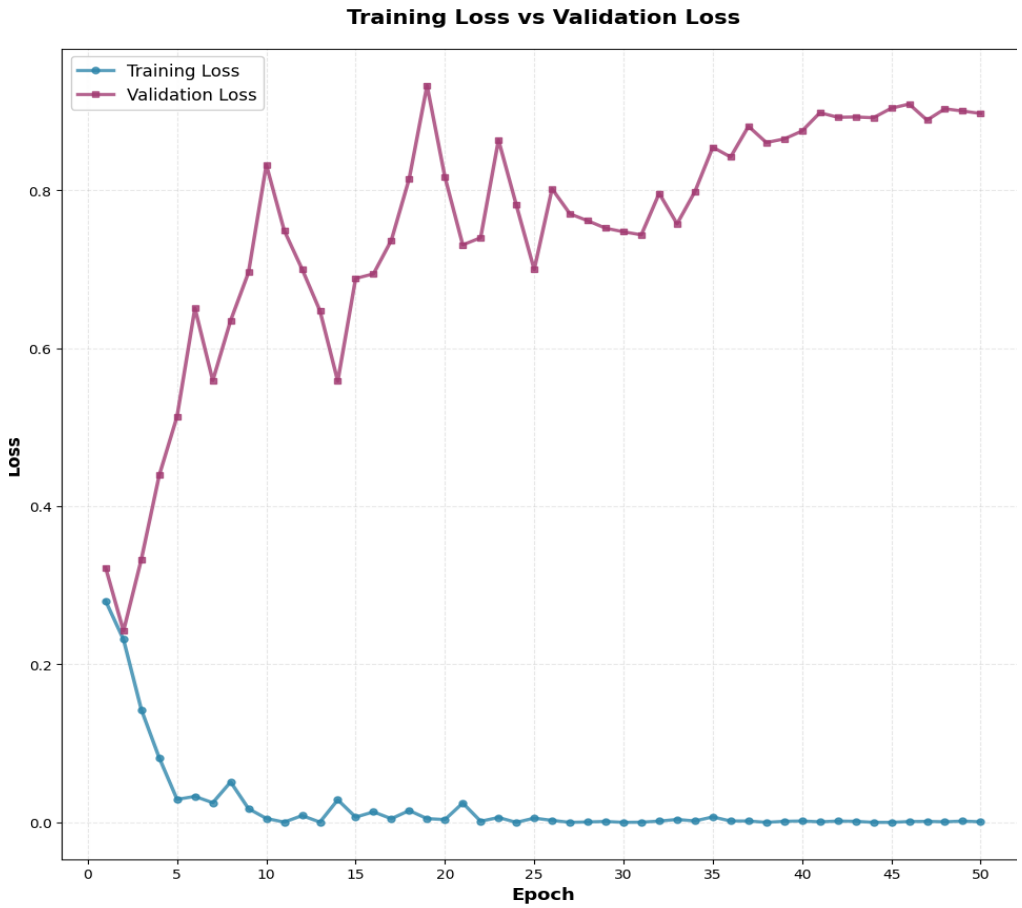
Skenario pertama



Skenario pertama menunjukkan konvergensi yang stabil dengan training loss yang menurun tajam pada epoch epoch awal dan mencapai nilai hamper nol di epoch 10

Hasil & Pembahasan

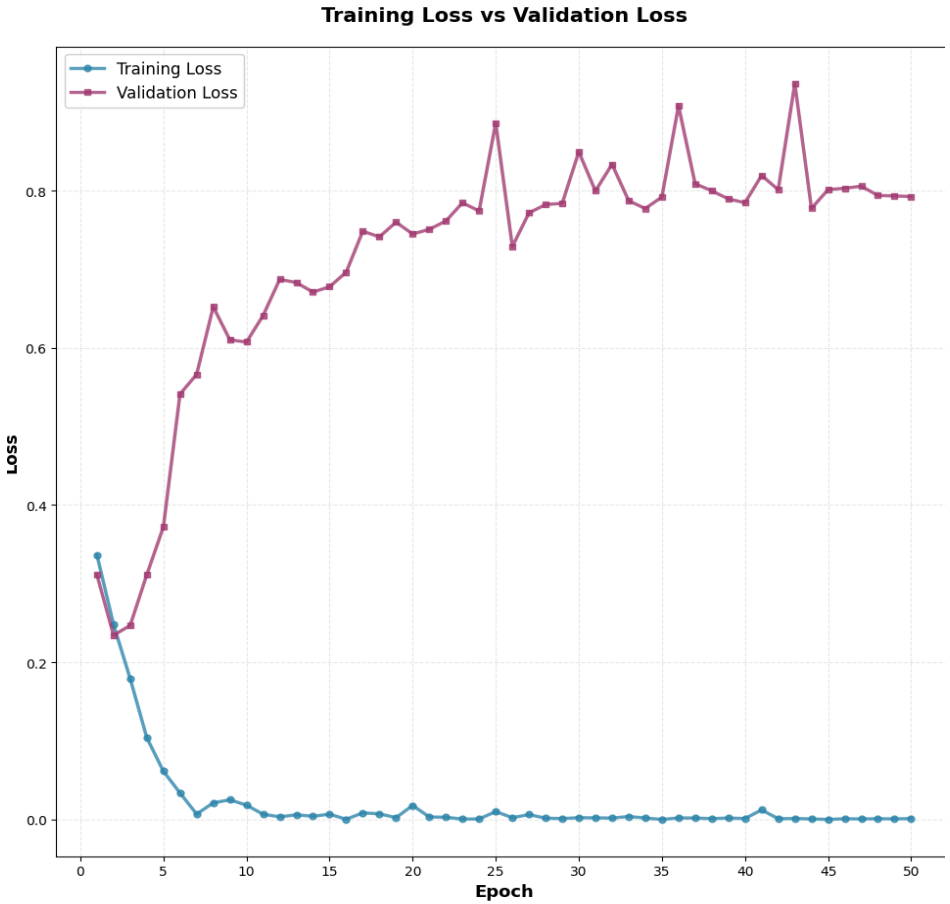
Skenario Kedua



Skenario kedua penurunan performa menurun sehingga mengkonfirmasi bahwa learning rate lebih tinggi menyebabkan overshooting.

Hasil & Pembahasan

Skenario Ketiga



Skenario ketiga menunjukan performa meningkat sehingga mengindikasi bahwa batch size lebih tinggi menhasilkan model yang lebih baik dalam mengidentifikasi kelas dengan benar.

Hasil & Pembahasan

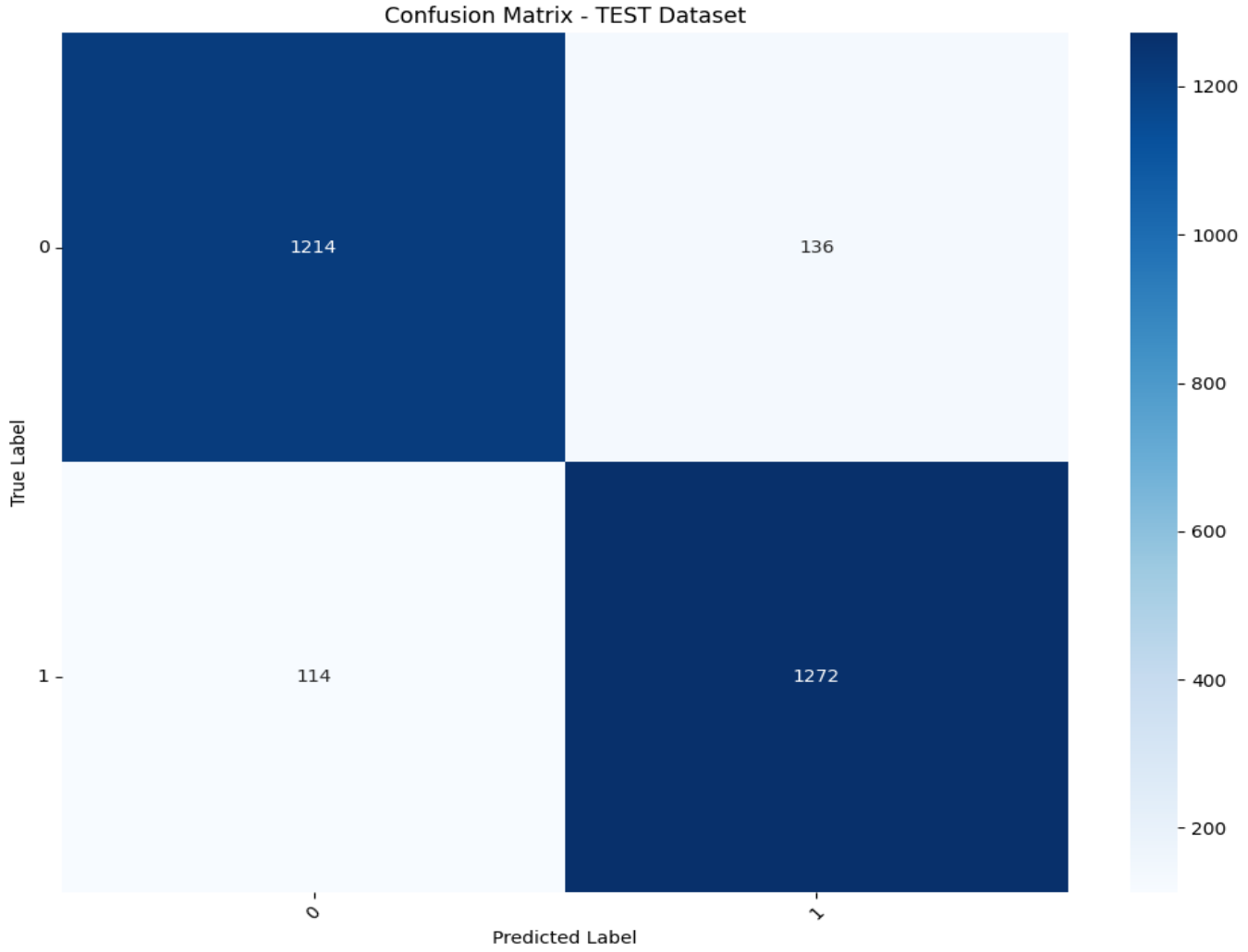
Skenario pertama

```
TEST RESULTS:
Accuracy: 0.9086
F1 Score: 0.9086

Classification Report (test):
              precision    recall  f1-score   support

           0       0.91      0.90      0.91      1350
           1       0.90      0.92      0.91      1386

    accuracy                           0.91      2736
   macro avg       0.91      0.91      0.91      2736
weighted avg       0.91      0.91      0.91      2736
```



Hasil & Pembahasan

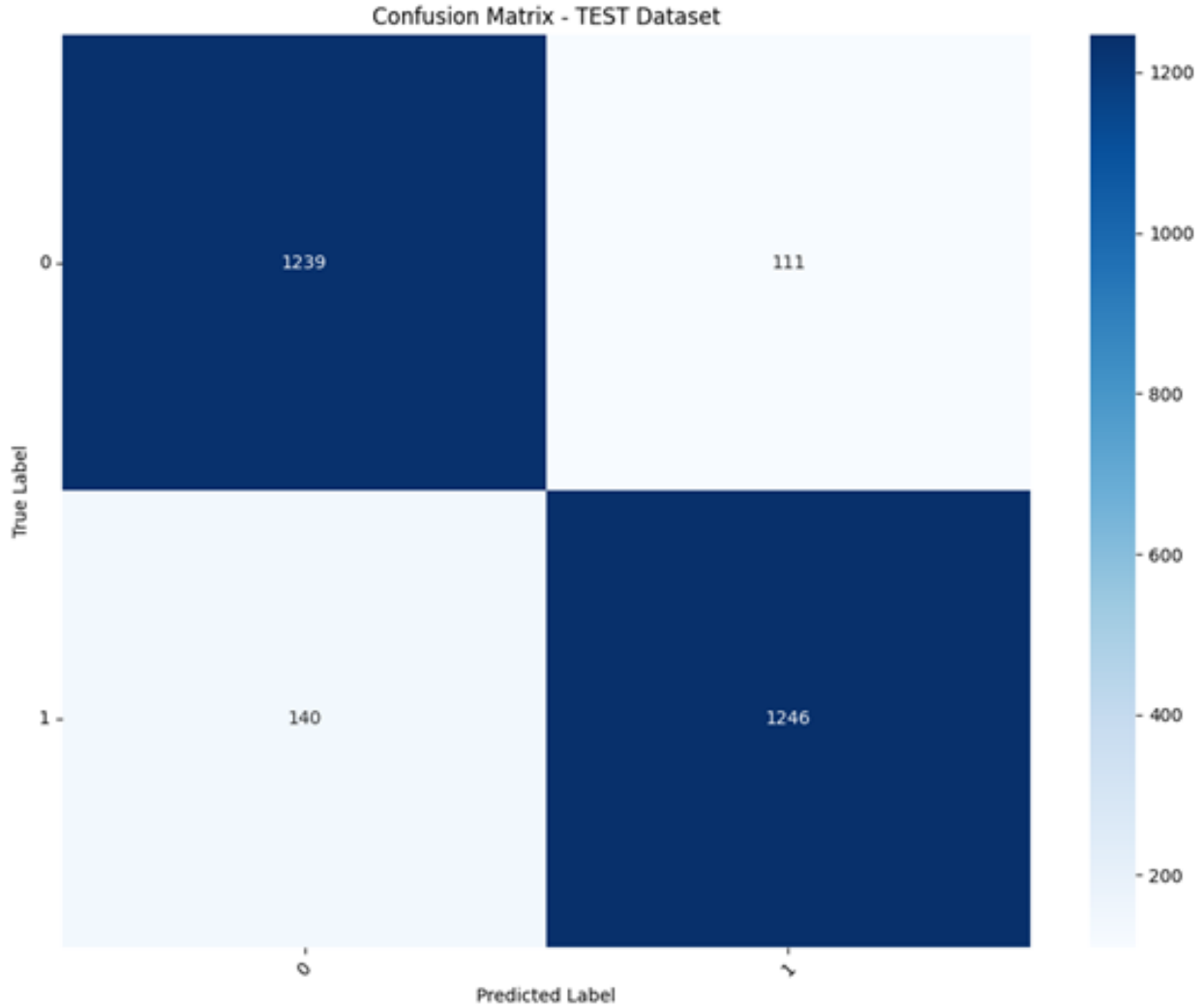
Skenario Kedua

```
TEST RESULTS:
Accuracy: 0.9083
F1 Score: 0.9083

Classification Report (test):
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      1350
           1       0.92      0.90      0.91      1386

    accuracy                           0.91      2736
   macro avg       0.91      0.91      0.91      2736
weighted avg       0.91      0.91      0.91      2736
```



Hasil & Pembahasan

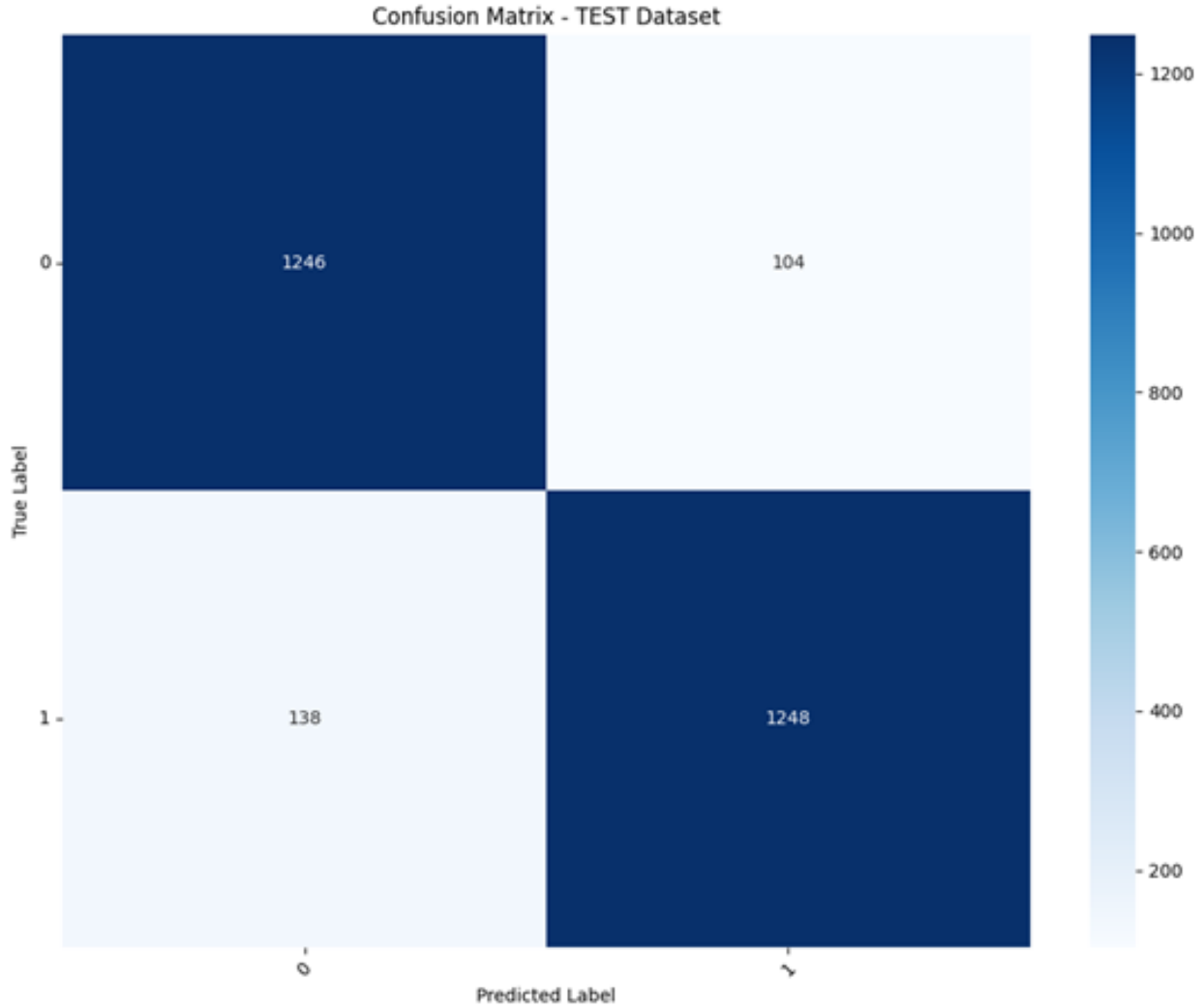
Skenario Ketiga

```
TEST RESULTS:
Accuracy: 0.9115
F1 Score: 0.9116

Classification Report (test):
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      1350
           1       0.92      0.90      0.91      1386

    accuracy                           0.91      2736
   macro avg       0.91      0.91      0.91      2736
weighted avg       0.91      0.91      0.91      2736
```



Hasil & Pembahasan

Tabel 3. Hasil Performa Dan Evaluasi Ekperimen Awal

	Accuracy	F1 Score
Skenario Pertama	90.86%	90.86%
Skenario Kedua	90.83%	90.83%
Skenario Ketiga	**91.15%**	**91.16%**

Kesimpulan: Penggunaan Batch Size 32 memberikan konvergensi paling stabil dan signifikan mengurangi *False Negatives* (lebih baik dalam mendeteksi kelas positif).

Hasil & Pembahasan

Tabel 4. Hasil Performa Dan Evaluasi Analisis Optimizer

Optimazer	Scheduler	Accuracy	Precision	Recall
Adam	Linear	90.86%	91%	91%
	Cosine	**91.19%**	**91.5%**	**91%**
AdamW	Linear	90.86%	91%	91%
	Cosine	**91.19%**	**91.5%**	**91%**
SGD	Linear	71.93%	72%	72%
	Cosine	71.97%	72%	72%

Kesimpulan Akhir: Konfigurasi AdamW dan Cosine Annealing dipilih untuk eksperimen selanjutnya karena performa maksimal dan fitur *weight decay* yang lebih efektif untuk regularisasi.

Hasil & Pembahasan

Tabel 5. Hasil Performa Dan Evaluasi Perbandingan Model

Model	Accuracy	Precision	Recall	F1-Score
IndoBERTweet + BiLSTM	**91.48%**	**91.48%**	**91.48%**	**91.48%**
IndoBERTweet + BiGRU	91.19%	91.22%	91.21%	91.19%
IndoBERTweet	91.19%	91%	91%	91.19%
CNN-LSTM Hybrid	88.49%	89%	88%	88%
BiLSTM	88.78%	88%	88%	88.78%
BiGRU	87.39%	87%	87%	87.39%
SVM + TF-IDF	78.18%	82.50%	78.42%	77.53%

Kesimpulan Akhir: Pendekatan deep learning berbasis pre-trained language models terbukti secara signifikan mengungguli metode machine learning tradisional dalam deteksi cyberbullying pada teks code-mixed. Model IndoBERTweet - BiLSTM mencapai performa terbaik dengan akurasi 91.48%, unggul 13.3% dibandingkan SVM dan TF-IDF, sehingga lebih andal dalam memahami konteks dan makna implisit.

Referensi

- Agustin, D. C., Rosid, M. A., & Ariyanti, N. (2023). Implementasi Convolutional Neural Network Untuk Deteksi Kesegaran Pada Apel. Jurnal Fasilkom, 13(02), 145–150. https://doi.org/10.37859/jf.v13i02.5175Abdulloh, N. and Hidayatullah, A. F. (2021) 'Deteksi Cyberbullying pada Cuitan Media Sosial Twitter', Automata, Vol 1(1), pp. 1–5.
- Alfifa, A. N. et al. (2025) 'Media Sosial dan Pembentukan Opini Publik (Analisis Studi Kasus Echo Chamber Pada Interaksi Komentar di Akun Instagram @ Turnbackhoaxid Dalam Konteks Post-Truth)', Jurnal Penelitian Ilmu-Ilmu Sosial, 2(6), pp. 162–169. Available at: https://ojs.daarulhuda.or.id/index.php/Socius/article/view/1130.
- Andika, A. J., Kristian, Y. and Setiawan, E. I. (2023) 'Deteksi Komentar Cyberbullying Pada YouTube Dengan Metode Convolutional Neural Network – Long Short-Term Memory Network (CNN-LSTM)', Teknika, 12(3), pp. 183–188. doi: 10.34148/teknika.v12i3.677.
- Fauzan Baehaqi and Cahyono, N. (2024) 'Analisis Sentimen Terhadap Cyberbullying Pada Komentar Di Instagram Menggunakan Algoritma Naïve Bayes', Indonesian Journal of Computer Science, 13(1), pp. 1051–1063. doi: 10.33022/ijcs.v13i1.3301.
- Forry Kusuma, J. and Chowanda, A. (2023) 'INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter', 7(September), pp. 773–780. Available at: www.joiv.org/index.php/joiv
- Hasan, N. F. (2021) 'Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor', Journal of Smart System, 1(1), pp. 35–44. doi: 10.36728/jss.v1i1.1605.
- Imani, F. A., Kusmawati, A. and Amin, H. M. T. (2021) 'Pencegahan Kasus Cyberbullying Bagi Remaja Pengguna Sosial Media', KHIDMAT SOSIAL: Journal of Social Work and Social Services, 2(1), pp. 74–83. Available at: https://jurnal.umj.ac.id/index.php/khidmatsosial/article/view/10433
- Koto, F., Lau, J. H. and Baldwin, T. (2021) 'INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization', EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833
- Kumala, A. P. B. and Sukmawati, A. (2020) 'Dampak Cyberbullying Pada Remaja', Alauddin Scientific Journal of Nursing, 1(1), pp. 55–65. doi: 10.24252/asjn.v1i1.17648.
- Machmud, A., Wibisono, B. and Suryani, N. (2025) 'Analisis Sentimen Cyberbullying Pada Komentar X Menggunakan Metode Naïve Bayes', 5(1).
- Masbadi Hatullah Nurnaryo, R. et al. (2022) 'Deteksi Cyberbullying Pada Data Tweet Menggunakan Metode Random Forest Dan Seleksi Fitur Information Gain', Jurnal Simantec, 11(1), pp. 33–40. doi: 10.21107/simantec.v11i1.17256.

Referensi

- Rizki, M. F., Auliasari, K. and Primaswara Prasetya, R. (2021) 'Analisis Sentiment Cyberbullying Pada Sosial Media Twitter Menggunakan Metode Support Vector Machine', JATI (Jurnal Mahasiswa Teknik Informatika), 5(2), pp. 548–556. doi: 10.36040/jati.v5i2.3808.
- Rosid, M. A., Siahaan, D. and Saikhu, A. (2024) 'Sarcasm Detection in IndonesianEnglish Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi- 24 Directional GRU', IEEE Access, (July), pp. 137063–137079. doi: 10.1109/ACCESS.2024.3436107.
- Safitri, M. et al. (2025) 'DETEKSI CYBERBULLYING TWEET MENGGUNAKAN MACHINE', pp. 370–374
- Santoso, H., Putri, R. A. and Sahbandi, S. (2023) 'Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest', Jurnal Manajemen Informatika (JAMIKA), 13(1), pp. 62–72. doi: 10.34010/jamika.v13i1.9303.
- Triyana, R., Virgantara Putra, O. and Pradhana, F. R. (2022) 'Deteksi Cyberbullying Pada Tweet Berbahasa Inggris Dengan Metode Support Vector Machine', Seminar Nasional Hasil Penelitian & Pengabdian Masyarakat Bidang Ilmu Komputer, pp. 98–103.
- Widiyantoro, P. and Prasetyo, Y. D. (2025) 'Deteksi Cyberbullying pada Pemain Sepak Bola di Platform Media Sosial " X " Menggunakan Metode Long Short-Term Memory (LSTM)'.
- Zakaria, P. H., Nurjannah, D. and Nurrahmi, H. (2023) 'Misogyny Text Detection on Tiktok Social Media in Indonesian Using the Pre-trained Language Model IndoBERTweet', Jurnal Media Informatika Budidarma, 7(3), p. 1297. doi: 10.30865/mib.v7i3.6438.

Sekian,
Terimakasih


UMSIDA
DARI SINI PENCERAHAN BERSEMI