The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 5, October 2024
https://doi.org/10.33022/ijcs.v13i5.4401

Confident Learning pada IndoBERT: Peningkatan Kinerja Klasifikasi Sentimen

Daffa Al Akhdaan[1], Taufik Edy Sutanto[2], Muhaza Liebenlito[3]
daffa.akhdaan20@mhs.uinjkt.ac.id[1], taufik.sutanto@uinjkt.ac.id[2],
muhazaliebenlito@uinjkt.ac.id[3]
[1,2,3] UIN Syarif Hidayatullah Jakarta

Informasi Artikel

Diterima : 9 Sep 2024
Direvisi : 18 Sep 2024
Disetujui : 29 Okt 2024

Kata Kunci

IndoBERT, *Confident Learning*, Analisis Sentimen, Kualitas Label

Abstrak

Dalam era perkembangan kecerdasan buatan (AI) yang pesat, masalah ketidakpastian dalam label *dataset* telah menjadi salah satu tantangan utama yang mengancam keberlanjutan AI. Penelitian ini mengkaji peningkatan kinerja model IndoBERT dalam analisis sentimen bahasa Indonesia dengan mengintegrasikan metode *confident learning* (CL). IndoBERT, sebagai adaptasi BERT untuk bahasa Indonesia, menunjukkan kinerja yang baik namun dipengaruhi oleh ketidakpastian label. CL diterapkan untuk memperbaiki label yang salah dan meningkatkan akurasi model. Hasilnya, IndoBERT + CL menunjukkan peningkatan akurasi dari 85,15% menjadi 86,03%, dengan perbaikan pada presisi, recall, dan F1 score masing-masing menjadi 87,93%, 85,00%, dan 86,44%. Hasil *confusion matrix* juga menunjukkan bahwa IndoBERT + CL lebih akurat dalam mengidentifikasi label positif. Penelitian ini menegaskan pentingnya penerapan CL untuk meningkatkan kualitas label dan performa model NLP dalam analisis sentimen.

Keywords

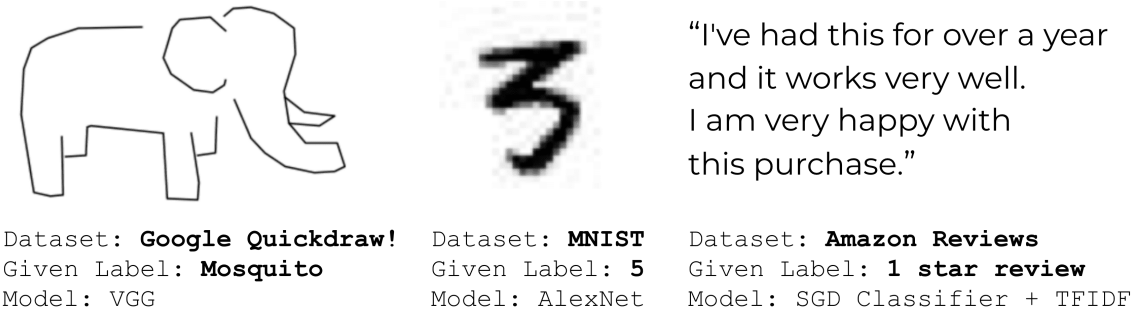
IndoBERT, Confident Learning, sentiment analysis, Label Quality

Abstract

In the rapidly evolving field of artificial intelligence (AI), label uncertainty in datasets has become a significant challenge threatening the sustainability of AI. This study investigates the enhancement of IndoBERT's performance in Indonesian sentiment analysis by integrating the confident learning (CL) method. IndoBERT, an adaptation of BERT for Indonesian, shows strong performance but is affected by label uncertainty. CL is applied to correct mislabeled data and improve model accuracy. The results indicate that IndoBERT + CL achieves an accuracy improvement from 85.15% to 86.03%, with enhancements in precision, recall, and F1 score to 87.93%, 85.00%, and 86.44%, respectively. The confusion matrix results also show that IndoBERT + CL is more accurate in identifying positive labels. This research highlights the importance of applying CL to enhance label quality and model performance in NLP sentiment analysis.

A. Pendahuluan

Dalam era perkembangan kecerdasan buatan (AI) yang pesat, masalah ketidakpastian dalam label dataset telah menjadi salah satu tantangan utama yang mengancam keberlanjutan AI. Ketidakpastian dalam label dataset dapat mengakibatkan model AI menghasilkan prediksi yang tidak akurat atau bahkan menyesatkan seperti pada Gambar 1 [1], yang pada gilirannya dapat berdampak negatif pada keberlanjutan dan keandalan sistem AI.



Gambar 1. Kesalahan Label pada Data, sumber gambar [1]

Dalam upaya mengatasi masalah ini, BERT (*Bidirectional Encoder Representations from Transformers*) adalah model NLP inovatif yang menggunakan pendekatan bidirectional untuk memahami konteks teks secara lebih mendalam. Dengan cara ini, BERT menangkap makna kata yang lebih kaya dengan mempertimbangkan konteks dari kedua arah. Model ini dilatih secara mendalam pada teks tanpa label dan kemudian dapat dikembangkan untuk tugas spesifik dengan menambahkan lapisan *output* tambahan [2]. BERT telah terbukti efektif dengan mencapai hasil terbaru pada berbagai tugas NLP, termasuk peningkatan signifikan pada skor GLUE [3] menjadi 80.5%, yaitu sebuah *benchmark* yang dirancang untuk mengukur kinerja model pemrosesan bahasa alami (NLP) pada berbagai tugas pemahaman bahasa dan akurasi MultiNLI menjadi 86.7% yaitu *dataset* yang dibuat untuk memahami kalimat dalam konteks pengembangan dan evaluasi model [4]. Keberhasilan BERT terletak pada kesederhanaan konsep dan kekuatan empirisnya dalam meningkatkan performa aplikasi NLP.

Menindaklanjuti inovasi ini IndoBERT [5] hadir sebagai model NLP yang dirancang khusus untuk bahasa Indonesia, memanfaatkan arsitektur BERT yang telah terbukti efektif. IndoBERT dilatih dengan dataset Indo4B, yang mencakup berbagai sumber teks seperti media sosial, berita, dan blog, untuk menangkap kekayaan bahasa Indonesia secara lebih akurat. Proses pelatihan IndoBERT dilakukan dalam dua fase: fase pertama dengan panjang urutan maksimum 128 token, dan fase kedua dengan panjang urutan maksimum 512 token. *Preprocessing* data pada IndoBERT meliputi pemisahan teks yang sesuai untuk menangani urutan panjang dan pengolahan morfologi bahasa Indonesia yang kompleks. Model ini menggunakan *tokenizer* SentencePiece dengan ukuran kosakata yang berbeda untuk versi standar dan versi ringan, serta dilatih menggunakan loss pemodelan bahasa tersembunyi. Pengaturan *batch size* dan *learning rate* disesuaikan untuk memastikan representasi kalimat yang akurat dan kontekstual, meningkatkan kinerja dalam berbagai aplikasi NLP dalam bahasa Indonesia.

IndoBERT-base, sebagai versi dasar dari IndoBERT, telah dilatih dengan korpus yang terdiri dari 5,5 miliar kata dari berbagai jenis teks bahasa Indonesia. Model ini dirancang untuk berbagai tugas NLP dan memiliki arsitektur yang terdiri dari 12 lapisan transformator, masing-masing dengan 12 heads, serta total 110 juta parameter. Struktur ini memungkinkan IndoBERT-base untuk menangani berbagai tugas pemrosesan bahasa dengan efektif, menjadikannya alat yang berharga dalam aplikasi NLP dalam bahasa Indonesia [6].

Sejalan dengan upaya peningkatan deteksi berita palsu, penelitian terbaru oleh Anugerah Simanjuntak et al. mengeksplorasi penggunaan IndoBERT-base-p1 dalam mendeteksi berita palsu dan meningkatkan kinerjanya melalui penyetelan *hyperparameter* dengan metode *Bayesian optimization*, *grid search*, dan *random search*. Penelitian ini menunjukkan bahwa *Bayesian optimization* adalah metode tuning yang paling efektif, mencapai presisi 88,79%, *recall* 94,5%, dan *F1-score* 91,56% untuk label "*fake*" [7]. Metode ini tidak hanya mengungguli metode *tuning* lainnya tetapi juga model dengan *hyperparameter default*, menegaskan pentingnya *hyperparameter tuning* dalam meningkatkan akurasi model deteksi berita palsu dan memberikan kontribusi yang berharga dalam melawan disinformasi di dunia digital.

Pendekatan *confident learning* merupakan salah satu solusi krusial dalam mengatasi masalah ketidakpastian dalam label *dataset*. Metode ini memungkinkan untuk meningkatkan estimasi ketidakpastian dalam label *dataset*, sehingga kualitas label yang digunakan untuk melatih model AI dapat ditingkatkan secara signifikan [1]. Dengan demikian, penggunaan *confident learning* bersama dengan teknik-teknik canggih seperti analisis sentimen menggunakan IndoBERT, model BERT yang diadaptasi untuk bahasa Indonesia, *confident learning* memperbaiki kualitas data pelatihan dan memperkuat kemampuan IndoBERT dalam memahami dan memproses teks bahasa Indonesia secara lebih akurat. Kombinasi ini meningkatkan keandalan dan efektivitas sistem AI dalam analisis sentimen dan pemrosesan bahasa alami.

Dalam *confident learning*, fokus utamanya adalah pada penggunaan tingkat keyakinan atau kepercayaan model terhadap prediksi-prediksinya untuk mengidentifikasi dan memperbaiki label-label yang salah dalam data latihan, metode *worst token* menggunakan tingkat keyakinan dari model klasifikasi token untuk menilai kualitas label-label pada tingkat token atau kalimat [8]. Dengan memanfaatkan informasi tingkat keyakinan ini, peneliti dapat mengidentifikasi kalimat-kalimat yang kemungkinan mengandung token dengan label yang salah. Pendekatan ini merupakan penerapan prinsip *confident learning* dalam konteks deteksi *label errors* dalam *dataset* klasifikasi token, yang bertujuan untuk meningkatkan kualitas data yang digunakan dalam berbagai aplikasi pemrosesan bahasa alami. Dengan mengadopsi pendekatan ini, diharapkan bahwa sistem AI dapat memperbaiki keakuratan analisis sentimen terhadap opini publik, seperti dalam konteks pemilihan presiden, yang krusial untuk mendukung proses demokrasi yang transparan dan akuntabel.

Pembeda dari penelitian-penelitian sebelumnya adalah bahwa tidak ada penelitian yang secara khusus mengkaji penggabungan penggunaan antara *confident learning* dengan model IndoBERT. Penelitian ini menawarkan inovasi dengan mengeksplorasi bagaimana integrasi *confident learning* dapat

meningkatkan akurasi dan keandalan model IndoBERT dalam konteks pemrosesan bahasa alami, yang merupakan area yang belum banyak diteliti.

B. Metode Penelitian

Pada bagian ini akan diberikan penjelasan secara rinci data yang digunakan pada penlitian ini, prapemrosesan data, klasifikasi model IndoBERT dan *confident learning* menggunakan cleanlab.

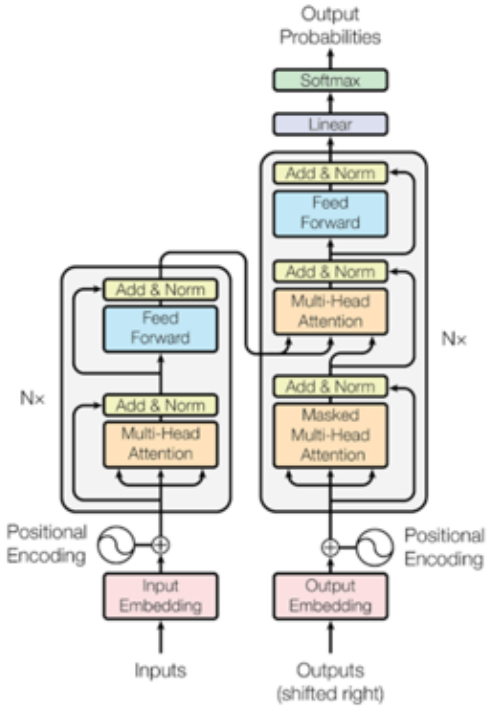
1. Data

Dalam penilitian ini, digunakan gabungan *dataset* sebanyak 2332 baris data yang terdiri dari analisis sentimen Pilkada DKI 2017 [9], analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler Indonesia [10], analisis sentimen *cyberbullying* [11], analisis sentimen terhadap tayangan televisi berdasarkan opini masyarakat [12], Analisis Sentimen Tentang Opini Film [13]. Data terdiri dari text dan label yang diklasifikasikan menjadi: positif dan negatif.

Karakter khusus, spasi berlebih, tag pengguna (*username*), tautan, *retweet*, *hashtag*, dan elemen lain yang dapat memengaruhi kinerja secara negatif telah dihapus, bersama dengan *stopwords*, nilai NaN, dan data yang duplikat. Label sentimen juga diganti, dengan negatif diubah menjadi 0 dan positif menjadi 1.

2. IndoBERT

Pelatihan model IndoBERT dimulai dengan penggunaan model IndoBERT yang telah dipra-latih sebelumnya. Untuk penelitian ini, model yang dipilih adalah IndoBERT-base-p1, yang dipilih karena keterbatasan ukuran *dataset*. IndoBERT-base, dengan lapisan dan parameter yang lebih sedikit, dianggap lebih efisien dalam pelatihan.



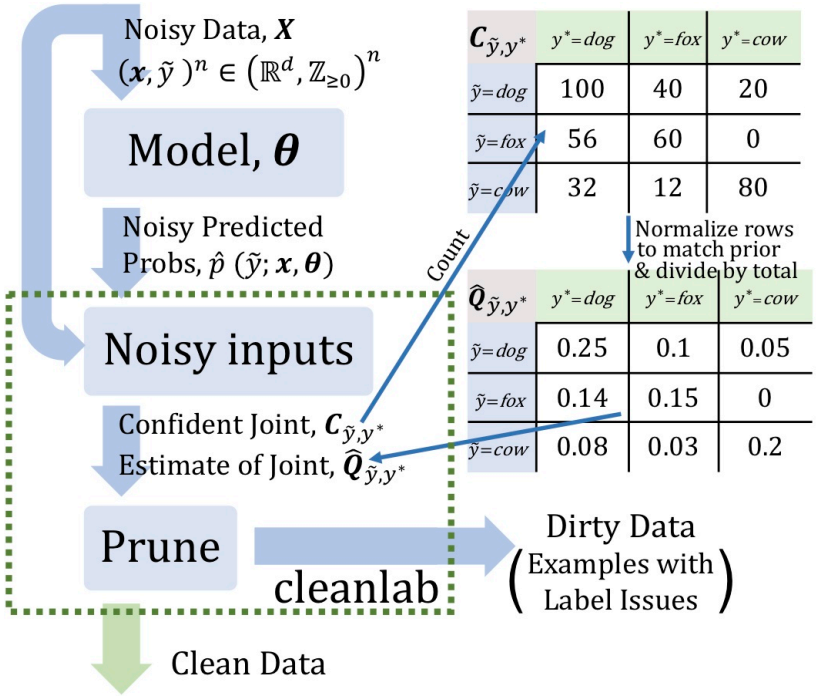
Gambar 2. Arsitektur Transformer[14]

Selama pelatihan, arsitektur model IndoBERT mengikuti arsitektur BERT, yang pada dasarnya menggunakan arsitektur Transformer seperti pada Gambar 2. Dalam konteks tugas spesifik seperti klasifikasi teks, data teks dimasukkan ke dalam model IndoBERT dengan panjang urutan maksimum 512 token.Teks yang melebihi panjang ini akan dipotong atau dibagi menjadi bagian yang lebih kecil agar sesuai dengan batasan [4].

Selama proses pelatihan, nilai hyperparameter yang telah ditetapkan dari penelitian sebelumnya digunakan untuk tiga parameter utama: lr (*learning rate)*, ukuran *batch*, dan jumlah *epoch*. Setelah pelatihan selesai, akurasi model IndoBERT akan dihitung dan dibandingkan dengan akurasi model IndoBERT yang menggunakan pendekatan *confident learning* untuk dicari label yang bermasalah.

3. Confident Learning

Confident Learning (CL) adalah metode yang dirancang untuk menangani masalah label yang salah dalam *dataset*. Metode ini bekerja dengan menggunakan dua input utama: probabilitas prediksi yang dihasilkan oleh model (misalnya, *Neural Network*, *Naive Bayes*, atau Regresi Logistik) dan vektor label yang bermasalah. Model yang digunakan untuk menghasilkan probabilitas prediksi ini tidak diubah selama proses CL. Sebaliknya, CL menggunakan output dari model yang sudah ada untuk mengidentifikasi dan memperbaiki label yang salah. Prosedur CL terdiri dari tiga langkah utama. Pertama, CL memperkirakan distribusi noise label untuk memahami ketidaksesuaian antara label yang diamati dan label yang sebenarnya. Kedua, CL menyaring contoh dengan label yang salah dari *dataset*, yang membantu dalam meningkatkan kualitas data. Ketiga, CL melatih model dengan data yang telah dibersihkan, dengan memberikan bobot kelas yang sesuai untuk memperbaiki pengaruh label yang salah. Meskipun CL tidak mengubah model secara langsung, kualitas probabilitas prediksi yang diberikan oleh model mempengaruhi efektivitas CL. Dengan menggunakan CL, proses pelatihan model menjadi lebih tahan terhadap label yang salah, memungkinkan model untuk belajar dari data yang lebih bersih dan akurat [7].



Gambar 3. Alur Confident Learning [1]

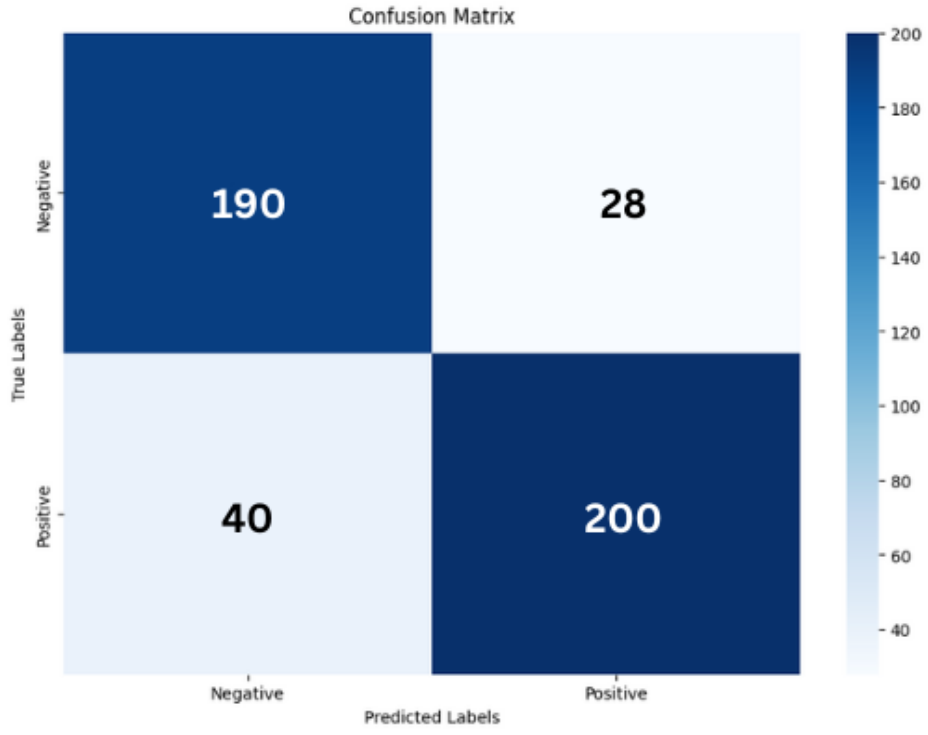
C. Hasil dan Pembahasan

Sebelumnya, model telah dilatih dan nilai *hyperparameter* dan konfigurasi utama diatur sebagai berikut: *criterion* diatur sebagai *torch.nn.CrossEntropyLoss*, fungsi loss yang digunakan untuk klasifikasi multi-kelas, membantu model dalam mengukur perbedaan antara prediksi dan label target. *optimizer* adalah *torch.optim.Adam*, yang digunakan untuk pembaruan bobot model, menawarkan efisiensi dalam pelatihan dengan adaptasi learning rate. Model dilatih selama *max_epochs* yaitu 10 epoch, menentukan berapa kali model akan melihat seluruh *dataset* pelatihan. *lr* atau *learning rate* ditetapkan pada 2e-5, yang mengontrol ukuran langkah dalam pembaruan bobot. *batch_size* adalah 32, menentukan jumlah sampel yang diproses dalam setiap batch pelatihan, mempengaruhi stabilitas dan kecepatan pelatihan.

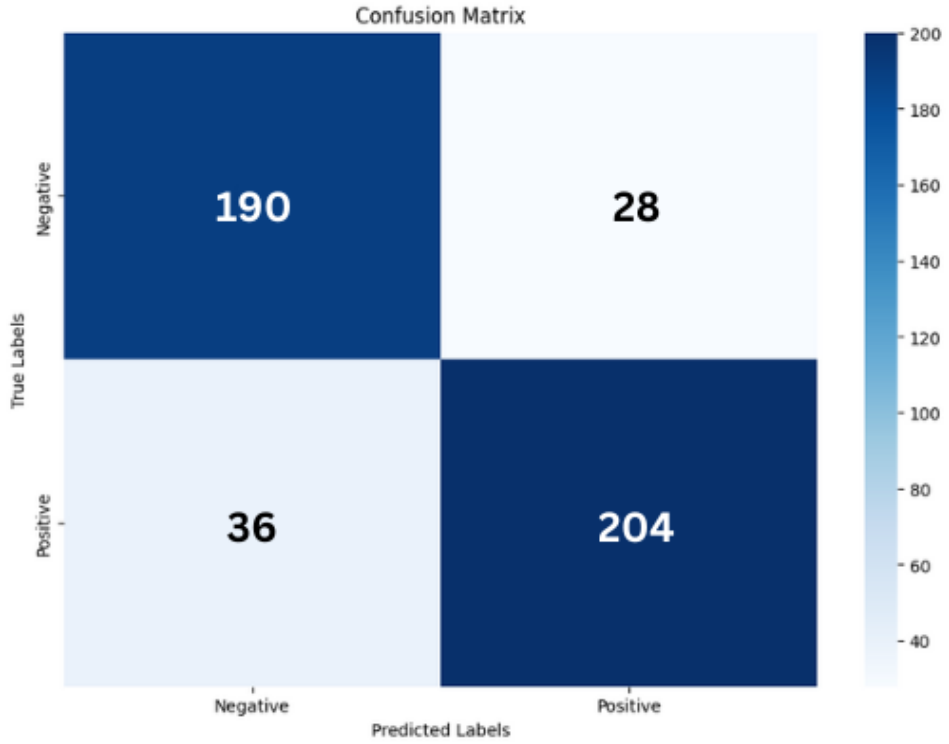
Dengan nilai *hyperparameter* yang dipilih kinerja model Indobert dan IndoBERT + CL menghasilkan perbedaan hasil *confusion matrix*, Gambar 4 dan Gambar 5 memperlihatkan hasil *confusion matrix* dari model IndoBERT dan IndoBERT + CL, terlihat bahwa IndoBERT memprediksikan dengan benar label negatif sebanyak 190 kali, yang menunjukkan jumlah label negatif yang benar-benar diprediksi sebagai negatif oleh model. Namun, ada 28 label negatif yang salah diklasifikasikan sebagai positif, yang tercermin dari nilai *False Positive* (FP). Di sisi lain, model memprediksi label positif dengan benar sebanyak 200 kali, yang menunjukkan jumlah label positif yang benar-benar diprediksi sebagai positif. Meskipun demikian, ada 40 label positif yang salah diklasifikasikan sebagai negatif, yang tercermin dari nilai *False Negative* (FN). Sementara itu, untuk hasil *confusion matrix* IndoBERT + CL, terlihat bahwa model memprediksikan dengan benar label negatif sebanyak 190 kali, yang menunjukkan jumlah label negatif yang benar-benar diprediksi sebagai negatif. Namun, terdapat 28 label negatif yang salah

diklasifikasikan sebagai positif, tercermin dari nilai *False Positive* (FP). Di sisi lain, model memprediksi label positif dengan benar sebanyak 204 kali, menunjukkan jumlah label positif yang berhasil diklasifikasikan dengan tepat. Meskipun demikian, ada 36 label positif yang salah diklasifikasikan sebagai negatif, yang tercermin dari nilai *False Negative* (FN). Secara keseluruhan, IndoBERT + CL dapat memprediksi label positif dengan lebih akurat dibandingkan dengan hanya menggunakan IndoBERT.

Hasil evaluasi model yang tertera pada Tabel 1 menunjukkan bahwa IndoBERT mencapai akurasi sebesar 0.8515, dengan presisi sebesar 0.8772, *recall* sebesar 0.8333, dan *F1 score* sebesar 0.8547. Sementara itu, IndoBERT + CL menunjukkan sedikit peningkatan dalam kinerja dengan akurasi sebesar 0.8603. Model ini juga mencatat presisi yang lebih tinggi yaitu 0.8793, serta *recall* yang meningkat menjadi 0.8500, dan *F1 score* yang lebih baik sebesar 0.8644. Peningkatan ini menunjukkan bahwa penambahan CL pada model IndoBERT memberikan perbaikan yang signifikan dalam hal kemampuan model untuk mengidentifikasi label positif secara lebih tepat (presisi) dan mengingat lebih banyak label positif yang relevan (*recall*), serta mencapai keseimbangan yang lebih baik antara presisi dan *recall*. Dengan demikian, IndoBERT + CL menunjukkan peningkatan performa yang konsisten dalam evaluasi metrik dibandingkan dengan model IndoBERT tanpa CL.



Gambar 4. *Confusion Matrix* IndoBERT



Gambar 5. *Confusion Matrix* IndoBERT + CL

Tabel 1. Perbandingan Evaluasi Model

Model	Akurasi	Presisi	Recall	F1 Score
IndoBERT	0.8515	0.8772	0.8333	0.8547
IndoBERT + CL	0.8603	0.8793	0.8500	0.8644

D. Simpulan

Penelitian ini menunjukkan bahwa integrasi *confident learning* (CL) dengan model IndoBERT memberikan peningkatan signifikan dalam kinerja klasifikasi sentimen bahasa Indonesia dibandingkan dengan penggunaan IndoBERT tanpa CL. Hasil evaluasi mengungkapkan bahwa IndoBERT + CL tidak hanya meningkatkan akurasi model dari 85.15% menjadi 86.03%, tetapi juga memperbaiki presisi, *recall*, dan F1 score, masing-masing menjadi 87.93%, 85.00%, dan 86.44%. Peningkatan ini mencerminkan kemampuan CL dalam mengatasi ketidakpastian label dengan memperbaiki label yang salah dan meningkatkan kualitas data pelatihan.

Confusion matrix menunjukkan bahwa IndoBERT + CL berhasil mengidentifikasi label positif dengan lebih akurat (204 prediksi benar) dibandingkan IndoBERT (200 prediksi benar), meskipun terdapat sedikit peningkatan pada label negatif yang salah diklasifikasikan. Secara keseluruhan, metode CL terbukti efektif dalam memperbaiki keandalan model, memberikan kontribusi penting dalam mengatasi tantangan ketidakpastian label dan meningkatkan performa sistem AI dalam analisis sentimen. Penelitian ini menegaskan pentingnya penggunaan teknik-teknik canggih seperti CL untuk meningkatkan akurasi dan keandalan model NLP dalam konteks bahasa Indonesia.

E. Referensi

[1] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," 2022. [Online]. Available: https://arxiv.org/abs/1911.00068

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[3] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," 2019. [Online]. Available: https://arxiv.org/abs/1804.07461

[4] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," *CoRR*, vol. abs/1704.05426, 2017, [Online]. Available: http://arxiv.org/abs/1704.05426

[5] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.85

[6] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *CoRR*, vol. abs/2011.00677, 2020, [Online]. Available: https://arxiv.org/abs/2011.00677

[7] Anugerah Simanjuntak *et al.*, "Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 1, pp. 60–67, Feb. 2024, doi: 10.22146/jnteti.v13i1.8532.

[8] W.-C. Wang and J. Mueller, "Detecting Label Errors in Token Classification Data," 2022. [Online]. Available: https://arxiv.org/abs/2210.03920

[9] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayes dan Pembobotan Emoji," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[10] U. Rofiqoh, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[11] W. Athira Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[12] W. E. Nurjanah, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media

Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[13] P. Antinasari, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[14] A. Vaswani *et al.*, "Attention Is All You Need," *CoRR*, vol. abs/1706.03762, 2017, [Online]. Available: http://arxiv.org/abs/1706.03762