Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.3, June 2025, pp. 989~997
e-ISSN: 2548-6861

Enhancing Medical Named Entity Recognition with Ensemble Voting of BERT-Based Models on BC5CDR

Fadhli Faqih Maulana [1]*, Abu Salam [2]**
* Teknik Informatika, Universitas Dian Nuswantoro
111202113473@mhs.dinus.ac.id [1], abu.salam@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-04-29
Revised 2025-06-02
Accepted 2025-06-17

Keyword:

BC5CDR,
BioBERT,
TinyBERT,
ClinicalBERT,
Ensemble Voting.

ABSTRACT

The rapid development in biotechnology and medical research has resulted in a large amount of scientific literature containing critical information about various medical entities. However, the primary challenge in managing this data is the vast volume of unstructured text, which requires *Natural Language Processing* (NLP) techniques for automatic information extraction. One of the main applications in NLP is *Named Entity Recognition* (NER), which aims to identify important entities in the text, such as disease names, drugs, and proteins. This study aims to enhance the performance of medical *Named Entity Recognition* (NER) by applying ensemble Voting to three BERT-based models: *BioBERT*, *TinyBERT*, and *ClinicalBERT*. The results show that the ensemble voting technique provides the best performance in medical entity extraction, with improvements in precision (0.9494), recall (0.9483), and F1-score (0.9488) compared to individual models, especially when handling less common medical entities. This approach is expected to contribute to the development of automated systems for analyzing and searching information in medical literature.

I. PENDAHULUAN

Perkembangan pesat dalam bidang bioteknologi dan penelitian medis telah menghasilkan sejumlah besar literatur ilmiah yang memuat informasi penting tentang berbagai entitas medis. Namun, tantangan utama dalam mengelola data ini adalah volume besar teks yang tidak terstruktur , yang memerlukan teknik pemrosesan bahasa alami (*Natural Language Processing*/NLP) untuk mengekstraksi informasi secara otomatis. Salah satu aplikasi utama dalam NLP adalah *Named Entity Recognition* (NER), yang bertujuan untuk mengidentifikasi entitas penting dalam teks, seperti nama penyakit, obat, atau protein[1]. NER merupakan langkah pertama yang penting dalam pembuatan sistem pengetahuan berbasis teks yang dapat digunakan untuk mendalami hubungan antar-entitas medis, sehingga memungkinkan pengembangan basis data medis yang lebih terstruktur dan dapat di akses dengan lebih mudah [2]

Seiring dengan kemajuan teknologi, pendekatan berbasis model pembelajaran mendalam (deep learning) seperti BERT (Bidirectional Encoder Representations from Tranformers) telah menunjukkan performa luar biasa dalam berbagai tugas NLP, termasuk NER. Model BERT, khusus nya varian yang diadaptasi untuk domain medis seperti *BioBERT*, telah berhasil meningkatkan hasil dalam berbagai aplikasi teks medis, berkat kemampuannya untuk memahami konteks kata dalam teks dengan cara yang lebih baik daripada model sebelumnya. Namun, meskipun model-model ini telah mencapai kemajuan signifikan, mereka masih menghadapi tantangan dalam hal generalisasi antar berbagai jenis entitas dan hubungan yang lebih kompleks [3].

Penerapan teknik NER dalam teks medis telah banyak digunakan untuk berbagai tujuan, seperti ekstraksi hubungan antara obat dan penyakit, serta deteksi interaksi antar protein. Salah satu model yang paling banyak digunakan dalam tugas NER untuk teks biomedis adalah *BioBERT*, yang merupakan varian dari model BERT yang telah dilatih khusus dengan teks medis dari PubMed dan literatur terkait, *BioBERT* menunjukkan kinerja yang lebih baik di bandingkan dengan model BERT biasa dalam tugas-tugas NER pada domain biomedis [2]. Namun, meskipun model ini efektif dalam banyak kasus, masih ada ruang untuk meningkatkan akurasi dalam ekstraksi entitas yang lebih kompleks dan hubungan

antar-entitas, terutama dengan teks yang tidak terstruktur dan memiliki variasi gaya penulisan yang tinggi.

Model lain yang semakin mendapat perhatian adalah *TinyBERT*, yang merupakan versi lebih ringan dari BERT dengan ukuran parameter yang lebih kecil, sehingga dapat memproses data lebih cepat tanpa mengurangi akurasi secara signifikan. Meskipun *TinyBERT* lebih cepat dalam pengolahan data, namun performanya dalam ekstraksi entitas medis masih perlu di tingkatkan [4], [5], [6]. *TinyBERT*, meskipun merupakan model distilasi dari BERT yang lebih ringan dan tidak spesifik untuk domain medis, telah di-fine-tune pada dataset medis yang relevan dalam penelitian ini. Fine-tuning ini dilakukan dengan menggunakan dataset BC5CDR, yang berfokus pada entitas medis seperti penyakit dan bahan kimia. Proses fine-tuning memungkinkan *TinyBERT* untuk lebih efektif dalam menangani teks medis dan meningkatkan akurasi dalam tugas *Named Entity Recognition* (NER) di domain medis. Selain itu, *ClinicalBERT*, yang dirancang khusus untuk data medis dan klinis, juga menunjukkan hasil yang menjanjikan dalam pemrosesan teks medis. Dengan pendekatan domain-spesifik, *ClinicalBERT* mampu mengenali entitas medis yang lebih relevan dengan konteks klinis, namun terkadang kesulitan dalam menangani variasi bahasa yang lebih luas [7], [8], [9].

Dalam beberapa tahun terakhir, teknik ensemble voting telah diterapkan untuk menggabungkan kekuatan dari beberapa model secara keseluruhan. Pendekatan ini sangat efektif dalam mengatasi kelemahan model tunggal dengan mengkombinasikan prediksi dari beberapa model yang saling melengkapi. Sebagai contoh, penelitian oleh Jia et al. (2024) menunjukkan bahwa teknik ensemble berbasis stacking dapat meningkatkan kinerja ekstraksi hubungan dalam teks biomedis secara signifikan dengan menggunakan berbagai model BERT yang sudah di latih sebelumnya, seperti *BioBERT*, PubMedBERT, dan BlueBERT [10].

Pada penelitian ini, teknik ensemble voting digunakan untuk menggabungkan kekuatan dari tiga model BERT-based: *BioBERT*, *TinyBERT*, dan *ClinicalBERT*. Jenis voting yang digunakan adalah majority voting, di mana setiap model memberikan suara berdasarkan prediksinya, dan suara mayoritas menentukan label akhir untuk entitas yang terdeteksi[11], [12]. Proses ini memungkinkan penggabungan kekuatan masing-masing model untuk meningkatkan akurasi dalam ekstraksi entitas medis. Dataset BC5CDR, yang digunakan dalam penelitian ini, merupakan dataset yang dikembangkan dalam BioCreative V, yang berfokus pada ekstraksi hubungan antara penyakit dan obat. Dataset ini mencakup entitas medis penting seperti nama penyakit dan bahan kimia (obat), yang memungkinkan untuk analisis lebih lanjut mengenai interaksi antara penyakit dan obat dalam literatur medis. Pendekatan ini diharapkan dapat mengatasi keterbatasan model tunggal dalam mengenali entitas medis yang lebih kompleks. Tujuan utama dari penelitian ini adalah untuk meningkatkan akurasi ekstraksi entitas medis, dengan memanfaatkan kekuatan masing-masing model untuk menangani berbagai variasi dalam teks medis. Penelitian ini juga bertujuan untuk membandingkan kinerja model tunggal dengan model ensemble dalam konteks ekstraksi entitas pada dataset *BC5CDR*, yang berfokus pada hubungan penyakit dan obat [13], [14].

Manfaat yang diharapkan dari penelitian ini adalah peningkatan akurasi dalam ekstraksi entitas medis dari teks biomedis, yang dapat berkontribusi pada pengembangan sistem otomatis untuk pemrosesan dan analisis literatur medis. Dengan meningkatkan kinerja model NER, hasil penelitian ini dapat membantu mempercepat proses pencarian informasi dalam literatur medis, yang sangat penting untuk pengembangan obat, diagnosis penyakit, dan penelitian biomedis lainnya. Selain itu, hasil penelitian ini diharapkan dapat memberikan kontribusi bagi penerapan teknik ensemble dalam domain biomedis, serta memberikan wawasan tentang cara mengoptimalkan model berbasis BERT untuk tugas ekstraksi entitas medis [15].

II. METODE

A. Pengumpulan Data

Pada penelitian ini, data yang di gunakan berasal dari dataset *BC5CDR* yang mencakup informasi terkait hubungan penyakit dan obat, dan mencakup tiga dataset yaitu Test, Training dan Valid. Data tersebut merupakan data teks yang tidak terstruktur, yang diambil dari publikasi medis yang relevan dan berfokus pada ektraksi informasi medis. Dataset yang digunakan dalam penelitian ini mengandung entitas penting seperti nama penyakit dan obat. Data ini di kumpulkan melalui teknik web scraping dari sumber literatur medis untuk memastikan keberagaman dan kelengkapan data. Dataset masing-masing berjumlah sekitar 10.000 data medis dengan entitas yang berbeda-beda.

TABEL I
DISTRIBUSI LABEL

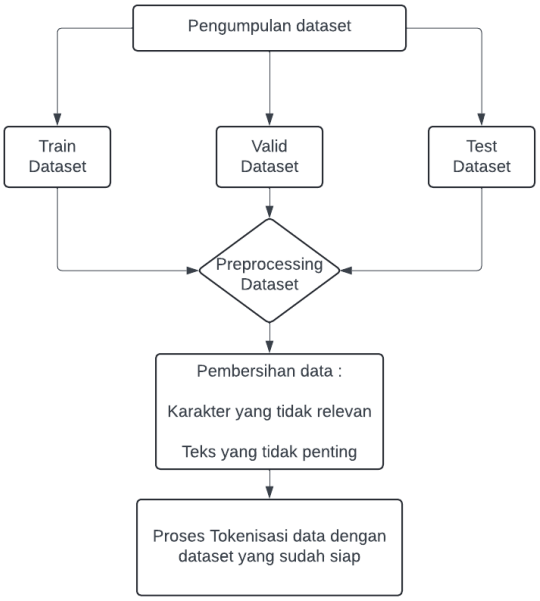
Label	Kode	Training	Validation	Test
O	0	99324	98681	106194
B-Chemical	1	5101	5270	5282
B-Disease	2	4033	4094	4321
I-Disease	3	2208	2119	2204
I-Chemical	4	519	484	354

Dataset yang digunakan dalam penelitian ini terdiri dari tiga bagian utama, yaitu Training, Validation, dan Testing, dengan jumlah entitas sebagai berikut: Training mencakup 110,185 entitas, Validation mencakup 110,648 entitas, dan Testing mencakup 118,355 entitas. Dataset ini berfokus pada ekstraksi entitas medis, khususnya untuk mengidentifikasi hubungan antara penyakit (disease) dan bahan kimia (chemical) dalam teks medis. Kedua jenis entitas tersebut diberi label sesuai dengan format anotasi BIO (Beginning, Inside, Outside), di mana entitas pertama dari jenis penyakit atau bahan kimia diberi label B-Disease atau B-Chemical, sementara bagian-bagian selanjutnya diberi label I-Disease atau I-Chemical, dan token yang tidak termasuk entitas diberi

label O [16]. Meskipun dataset ini terdiri dari tiga bagian, tidak ada pembagian manual dataset yang dilakukan dalam penelitian ini, dan data tersebut sudah terbagi berdasarkan format yang tersedia. Preprocessing yang dilakukan mencakup tokenisasi teks, penyaringan entitas, serta memastikan data yang ada sudah siap digunakan untuk pelatihan, validasi, dan pengujian model.

B. Preprocessing Data

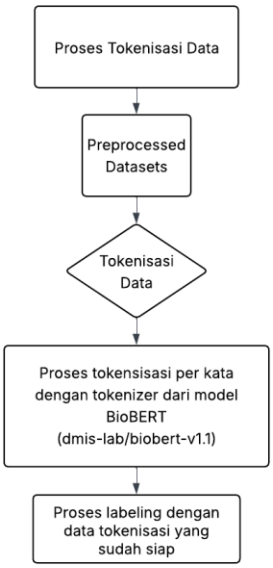
Sebelum memulai pemodelan, dilakukan beberapa tahap preprocessing untuk memastikan bahwa data yang digunakan dalam model sudah bersih dan siap untuk diproses lebih lanjut. Proses preprocessing dimulai dengan pembersihan data yang meliputi penghapusan karakter yang tidak relevan serta teks yang tidak penting. Setelah itu, data teks dibersihkan dari kata-kata atau karakter yang dapat mengganggu proses analisis entitas medis.



Gambar 1 Preprocessing Dataset

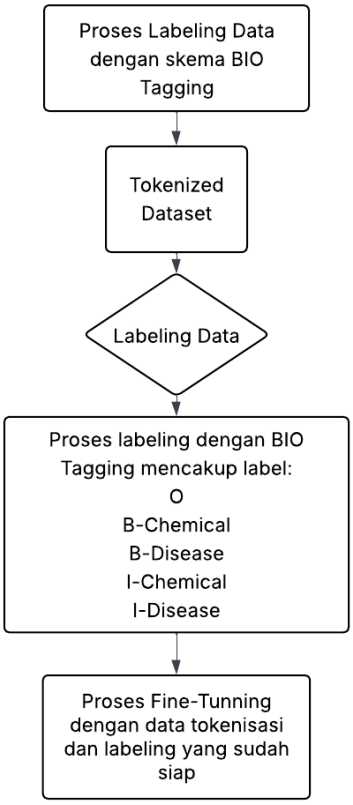
C. Tokenisasi Data

Tokenisasi adalah proses pemecahan teks menjadi potongan-potongan kecil yang disebut token, yang dapat berupa kata, frasa, atau karakter, tergantung pada tujuan pemodelan [17]. Dalam penelitian ini, tokenisasi dilakukan menggunakan tokenizer dari model *BioBERT* (dmis-lab/biobert-v1.1). Tokenisasi yang dilakukan oleh mode *BioBERT* memastikan bahwa kata-kata yang memiliki makna yang sama akan dipetakan ke dalam representasi vektor yang konsisten, memungkinkan model untuk menangkap hubungan konteks dalam teks dengan lebih baik.



Gambar 2 Proses Tokenisasi

D. Labeling Teks



Gambar 3 Proses Labeling

Proses labelling ini dilakukan berdasarkan kategori entitas yang telah ditentukan sebelumnya, seperti nama obat,

penyakit, dan faktor terkait lainya. Setiap entitas dalam teks akan diberikan label sesuai dengan klasifikasi yang berlaku. Misalnya, teks yang mengandung nama obat akan diberi label ”CHEMICAL”, sedangkan teks yang berisi penyakit akan diberi label ”DISEASE”. Dan juga labeling ini mencakup BIO Tagging (Beginning, Inside, Outside). Beginning mencakup nama awal entitas, Inside mencakup nama kedua dan setelahnya, dan Outside adalah entitas yang termasuk dalam label Chemical ataupun Disease

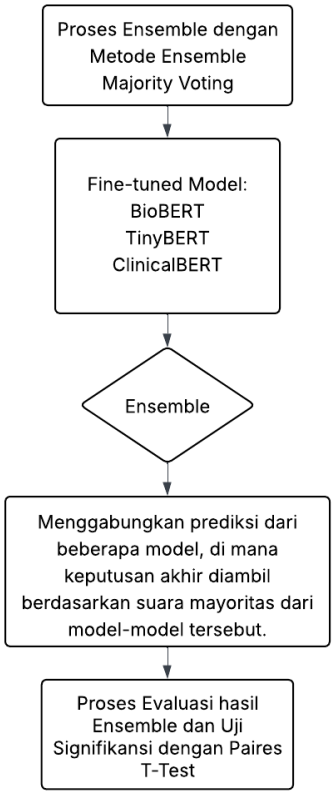
E. Pemodelan dan Fine Tuning Model BERT

Pada tahap ini, digunakan model BERT, khususnya *BioBERT*, *TinyBERT*, dan *ClinicalBERT* untuk proses ekstraksi entitas medis. Model BERT yang telah dilatih sebelumnya (pre-trained) digunakan sebagai dasar dan kemudian dilakukan fine-tuning dengan menggunakan dataset training yang telah di proses sebelumnya. Fine-tuning dilakukan untuk menyesuaikan model dengan domain medis, sehingga dapat, mengidentifikasi dan mengklasifikasikan entitas medis dengan lebih baik. Parameter pelatihan seperti Batch Size dengan ukuran 8 , learning rate 2e-5, dan jumlah optimasikan dengan optimizer adamw. Jumlah epoch 50 dan diatur dengan teknik Early stopping untuk menentukan titik jenuh pelatihan. Selama proses fine-tunning, model dilatih untuk mengenali pola-pola dalam data medis dan memperbaiki bobot model berdasarkan error yang terjadi selama pelatihan.

F. Ensemble Voting

Pada penelitian ini, Teknik ensemble voting digunakan untuk menggabungkan kekuatan dari tiga model NER yang berbeda: *BioBERT*, *TinyBERT*, dan *ClinicalBERT*. Teknik *ensemble voting* diterapkan untuk meningkatkan akurasi model dengan memanfaatkan prediksi dari ketiga model tersebut [18]. Setiap model memberikan suara berdasarkan prediksinya untuk masing-masing entitas yang di temukan dalam teks. Suara mayoritas digunakan untuk menentukan label akhir dari entitas tersebut.

Model *BioBERT* digunakan sebagai salah satu komponen dalam ensemble karena telah terbukti efektif dalam tugas-tugas NER pada domain medis[19]. Namun, *TinyBERT* dan *ClinicalBERT* ditambahkan untuk memperkuat generalisasi dan mengatasi berbagai kelemahan model tunggal. *TinyBERT*, yang lebih ringan dari BERT, dipilih untuk mengurangi waktu komputasi dan meningkatkan pemrosesan. *ClinicalBERT*, yang dirancang khusus untuk teks medis, dipilih untuk menangani konteks klinis dengan lebih baik. Dengan menggabungkan ketiga model ini dalam sebuah *ensemble voting*, diharapkan model dapat mengatasi kelemahan masing-masing model tunggal dan menghasilkan prediksi yang lebih akurat dan robust.



Gambar 4 Ensemble Majority Voting

Output token-label dari tiap model dalam ensemble ini digabungkan menggunakan pendekatan majority voting, di mana setiap model memberikan prediksi untuk setiap token atau span, dan label yang dipilih adalah yang paling sering muncul di antara model-model yang digunakan. Dalam beberapa kasus, kami juga mempertimbangkan weighted voting, di mana model dengan akurasi lebih tinggi diberikan bobot lebih besar dalam proses voting. Setiap entitas, seperti "Aspirin", diprediksi dengan cara menggabungkan hasil prediksi untuk setiap token dalam span entitas tersebut, dengan memilih label yang paling banyak dipilih oleh model-model yang ada. Sebagai contoh, untuk entitas "Aspirin", model-model seperti *BioBERT*, *TinyBERT*, dan *ClinicalBERT* memberikan label yang serupa, sehingga label B-Chemical dipilih dalam proses voting.

G. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan tiga metrik utama, yaitu *Precision*, *Recall*, dan *F1-Score*, yang sangat relevan dalam tugas *Named Entity Recognition* (NER), terutama ketika menghadapi dataset yang tidak seimbang. *Precision* mengukur seberapa banyak entitas yang diprediksi sebagai relevan, yang benar-benar relevan dengan entitas yang ada [20], yang dapat dihitung dengan rumus berikut:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

Recall Mengukur seberapa banyak entitas relevan yang berhasil ditemukan oleh model dibandingkan dengan total entitas relevan yang ada [20], yang dapat di hitung dengan rumus berikut:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

F1-Score menggabungkan kedua metrik tersebut untuk memberikan gambaran menyeluruh tentang keseimbangan antara *Precision* dan *Recall* [20]. *F1-Score* dapat di hitung menggunakan rumus berikut:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Selain itu, evaluasi model juga menghitung Rata–rata (Mean) dengan menjumlahkan semua nilai metrik (*Precision*, *Recall*, *F1-Score*) untuk setiap model dan label, kemudian membaginya dengan jumlah total entri (jumlah model x jumlah label) [20]. Rata-rata ini dihitung dengan rumus berikut:

$$Rata - rata = \frac{\sum_{i=1}^{n} x_i}{n} \quad (4)$$

Di sisi lain, Standar Deviasi (Standar Deviation) digunakan untuk mengukur variasi nilai metrik dari rata-rata, mengkuadratkan selisih tersebut, dan kemudian mengambil akar kuadrat dari rata-rata kuadrat selisih tersebut[19]. Standar deviasi dapat dihitung dengan menggunakan rumus berikut:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad (5)$$

Dimana x_i adalah nilai individu, μ adalah rata-rata, dan n adalah jumlah nilai.

III. HASIL DAN PEMBAHASAN

Penelitian ini mengoptimalkan model *Named Entity Recognition* (NER) dengan menggunakan pendekatan ensemble voting yang menggabungkan tiga model *BioBERT*, *TinyBERT*, dan *ClinicalBERT*. Model-model ini dilatih dengan data teks medis yang telah melalui serangkaian proses preprocessing, labeling, dan tokenisasi. Hasil dari proses pelatihan dan evaluasi mencakup metrik kinerja seperti *precision*, *recall*, *F1-score*, serta confusion matrix untuk mengukur efektifitas ekstraksi entitas medis. Pada bagian ini, hasil yang diperoleh dari setiap model akan dibandingkan, dan pembahasan mengenai kelebihan dan kekurangan masing-masing model serta penerapan teknik *ensemble voting* akan di sampaikan.

A. Hasil Pelatihan Model

Pada penelitian ini, model *Named Entity Recognition* (NER) diterapkan dengan menggunakan tiga model utama, yaitu *BioBERT*, *TinyBERT*, dan *ClinicalBERT*. Ketiga model ini dilatih menggunakan dataset *BC5CDR*, yang berfokus pada entitas medis terkait penyakit dan obat. Pelatihan dilakukan menggunakan teknik fine-tuning, dan hasil yang diperoleh mencakup training loss, validation loss, serta *precision*, *recall*, dan *F1-Score*.

Tabel berikut menunjukkan training loss, validation loss, dan *F1-Score* yang diperoleh selama proses pelatihan untuk masing-masing model:

TABEL II
HASIL PELATIHAN MODEL DI EPOCH TERAKHIR

Model dan Epoch	Train Loss	Val Loss	Precision	Recall	F1-Score
BioBERT (35)	0.003	0.354	0.8431	0.9048	0.8713
Tiny BERT(48)	0.074	0.244	0.8245	0.8459	0.8344
Clinical BERT (48)	0.001	0.388	0.8421	0.8933	0.8656

Hasil dari pelatihan selama 50 epoch dengan tambahan teknik Early Stopping untuk masing-masing model menunjukan penurunan yang signifikan pada training loss dan validation loss, yang menunjukkan kinerja terbaik dengan *precision*, *recall*, dan *F1-score* yang sangat baik pada entitas medis. *TinyBERT*, meskipun lebih ringan, sedikit tertinggal dalam hal *precision* dan *recall*, tetapi tetap memberikan hasil yang cukup baik. *ClinicalBERT* menunjukan hasil yang baik, terutama dalam konteks entitas medis yang lebih relevan dengan konteks klinis.

B. Evaluasi Metrik Model

Model-model yang dilatih kemudian dievaluasi menggunakan metrik *precision*, *recall*, *F1-Score*, dan accuracy pada data uji. Berikut adalah hasil evaluasi untuk masing-masing model:

TABEL III
EVALUASI MODEL METRIK

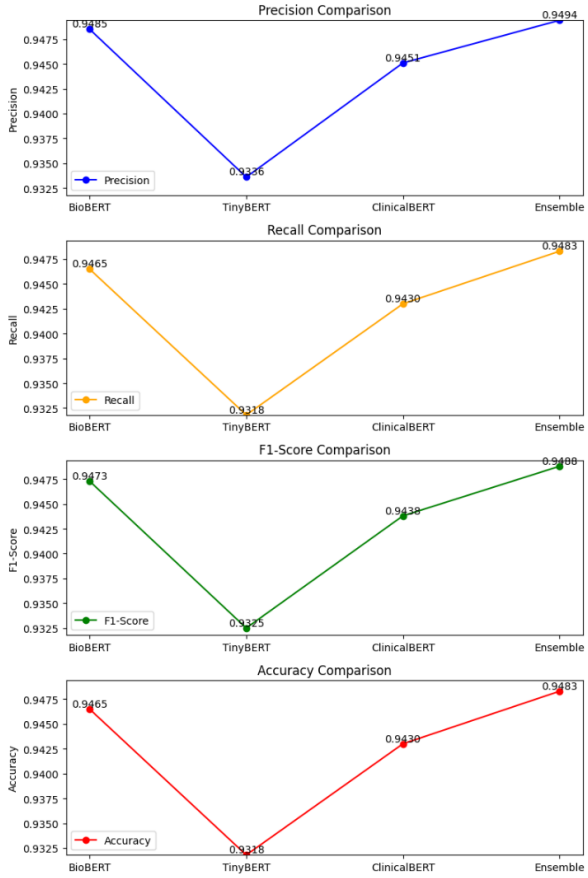
Model	Precision	Recall	F1-Score	Accuracy
BioBERT	0.9485	0.9465	0.9473	0.9465
TinyBERT	0.9336	0.9318	0.9325	0.9318
ClinicalBERT	0.9451	0.9430	0.9438	0.9430
Ensemble	0.9494	0.9483	0.9488	0.9483

Ensemble Voting yang menggabungkan prediksi dari *BioBERT*, *TinyBERT*, dan *ClinicalBERT* menunjukan hasil terbaik dengan *precision* 0.9494, *recall* 0.9483, dan *F1-score* 0.9488, yang lebih tinggi dibandingkan model tunggal. Hal ini menunjukkan bahwa teknik ensemble voting berhasil

meningkatkan kinerja model dalam tugas ekstraksi entitas medis.

C. *Perbandingan Kinerja Model*

Grafik berikut menggambarkan perbandingan *precision*, *recall*, dan *F1-score* untuk masing-masing model:



Gambar 5 Perbandingan hasil Metrik Antar Model

Grafik ini menunjukkan bahwa *BioBERT* dan *ClinicalBERT* memiliki kinerja yang sangat baik pada hampir semua metrik, namun *Ensemble Voting* memberikan hasil yang lebih konsisten dan robust, terutama pada *F1-score*. *TinyBERT* cenderung sedikit lebih rendah dalam hal *precision* dan *recall*, tetapi tetap menunjukkan kinerja yang baik dibandingkan dengan baseline yaitu *BioBERT*.

D. *Confusion Matrix dan Analisis Kesalahan*

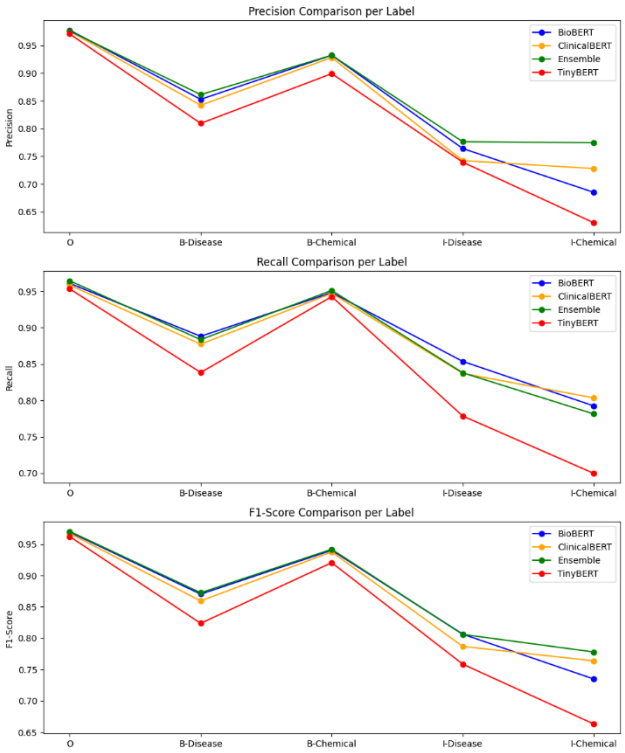
Analisis confusion matrix menunjukkan bahwa model *TinyBERT* dan *ClinicalBERT* lebih sering melakukan kesalahan pada entitas yang lebih langka, seperti I-Chemical dan I-Disease, dibandingkan dengan *BioBERT*. Model-model ini cenderung lebih kesulitan dalam mengenali entitas-entitas yang jarang ditemukan dalam data latih. Namun, dengan menggunakan *Ensemble Voting*, kesalahan klasifikasi pada entitas-entitas langka ini berkurang secara signifikan. Proses ensemble membantu mengurangi kesalahan prediksi, terutama dalam mengenali entitas yang kurang sering muncul dalam dataset pelatihan, dengan menggabungkan kekuatan dari berbagai model yang saling melengkapi.

E. *Visualisasi Hasil*

TABEL IV
HASIL PERLABEL

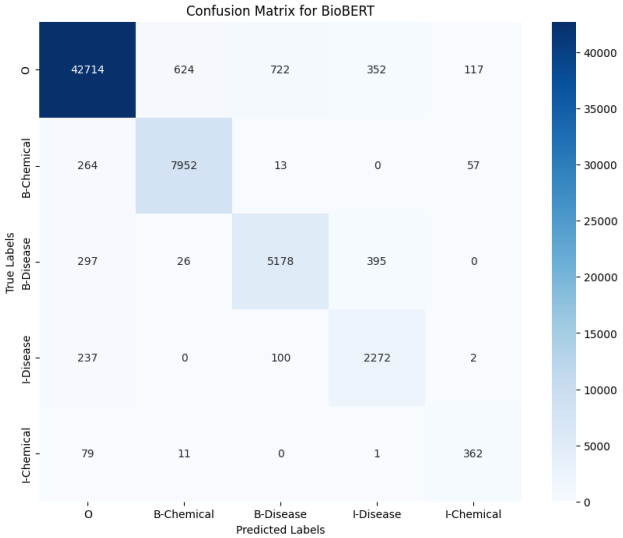
Label	Model	*Precision*	*Recall*	*F1-Score*
B-Chemical	*BioBERT*	**0.9324**	0.9482	0.9402
	TinyBERT	0.8992	0.9427	0.9204
	ClinicalBERT	0.9283	0.9471	0.9376
	Ensemble	0.9323	**0.9512**	**0.9417**
B-Disease	*BioBERT*	0.8529	**0.8882**	0.8702
	TinyBERT	0.8095	0.8384	0.8237
	ClinicalBERT	0.8421	0.8774	0.8594
	Ensemble	**0.8614**	0.8836	**0.8724**
I-Disease	*BioBERT*	0.7641	0.8537	**0.8064**
	TinyBERT	0.7395	0.7786	0.7586
	ClinicalBERT	0.7423	0.8372	0.7869
	Ensemble	**0.7764**	**0.8380**	0.8060
I-Chemical	*BioBERT*	0.6851	0.7925	0.7349
	TinyBERT	0.6302	0.6998	0.6632
	ClinicalBERT	0.7280	**0.8035**	0.7639
	Ensemble	**0.7746**	0.7815	**0.7780**
O	*BioBERT*	**0.9777**	0.9609	0.9692
	TinyBERT	0.9709	0.9534	0.9621
	ClinicalBERT	0.9759	0.9585	0.9671
	Ensemble	0.9762	**0.9645**	**0.9703**

Tabel IV menunjukkan perbandingan precision, recall, dan F1-score untuk masing-masing model dalam tugas *Named Entity Recognition* (NER), yaitu *BioBERT*, *TinyBERT, ClinicalBERT*, dan Ensemble. Hasil menunjukkan bahwa model Ensemble secara keseluruhan memberikan kinerja terbaik pada semua label entitas, dengan nilai recall tertinggi pada entitas B-Chemical (0.9512) dan B-Disease (0.8836), serta precision terbaik pada I-Chemical (0.7746) dan O (0.9762). Meskipun *TinyBERT* menunjukkan performa baik pada beberapa label, seperti B-Chemical, dan *ClinicalBERT* juga memberikan hasil solid pada B-Disease dan I-Disease, model Ensemble yang menggabungkan ketiga model ini berhasil mengurangi kesalahan klasifikasi, terutama pada entitas langka seperti I-Disease dan I-Chemical, dengan menghasilkan *F1-score* yang lebih tinggi secara keseluruhan. Hal ini menunjukkan bahwa penggunaan teknik Ensemble Voting mampu meningkatkan kinerja model dalam mengenali entitas medis, baik yang sering maupun jarang ditemukan dalam data pelatihan.

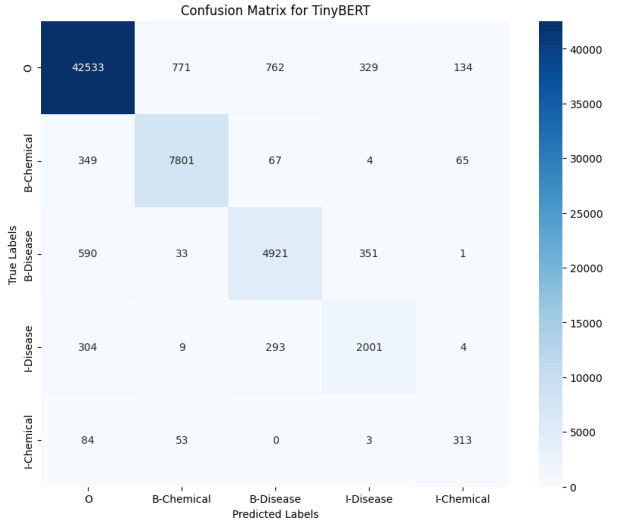


Gambar 6 Komparasi Hasil Perlabel

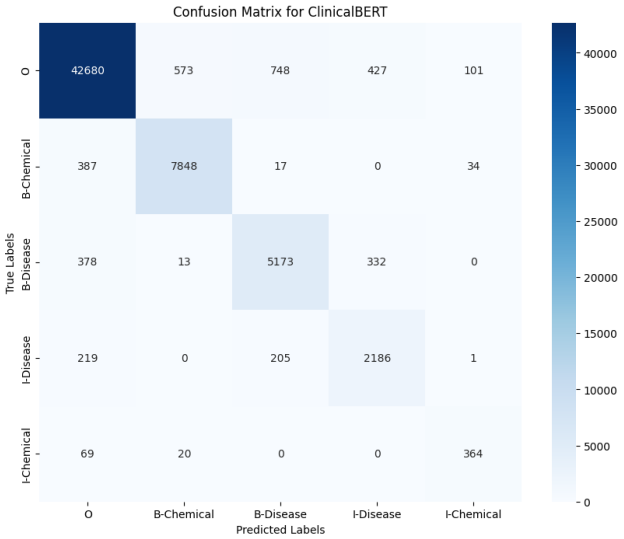
Dari grafik ini, dapat dilihat bahwa *Ensemble Voting* memberikan kinerja yang lebih baik dalam mengenali entitas I-Disease dan I-Chemical dibandingkan dengan model tunggal, seperti *TinyBERT*.



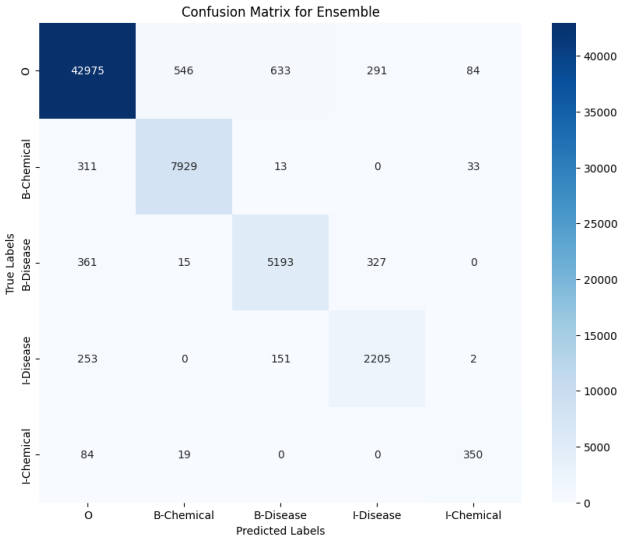
Gambar 7 Confusion Matrix BioBERT



Gambar 8 Confusion Matrix TinyBERT



Gambar 9 Confusion Matrix ClinicalBERT



Gambar 10 Confusion Matrix Ensemble Voting

Dalam penelitian ini, model Ensemble menunjukkan performa terbaik dalam klasifikasi entitas medis dibandingkan dengan model lain seperti *BioBERT, ClinicalBERT*, dan *TinyBERT*. Dengan menggabungkan beberapa model, Ensemble mampu mengurangi kelemahan masing-masing model dan menghasilkan prediksi yang lebih akurat, terutama dalam mengidentifikasi entitas yang lebih sulit, seperti I-Disease dan I-Chemical. Meskipun *BioBERT* dan *ClinicalBERT* menunjukkan hasil yang sangat baik, *Ensemble Voting* sedikit lebih unggul dalam hal akurasi dan kemampuan menangani kesalahan pada beberapa kelas, menjadikannya pilihan terbaik untuk tugas *Named Entity Recognition* (NER) dalam konteks ini.

F. Hasil Uji Statistik

Pada penelitian ini, T-Test digunakan untuk menguji perbedaan kinerja antara model-model yang diuji, yaitu BioBERT, TinyBERT, ClinicalBERT, dan Ensemble. Uji T-Test dilakukan pada tiga metrik evaluasi, yaitu Precision, Recall, dan F1-Score. Tabel di bawah ini menyajikan hasil perbandingan statistik antar model untuk ketiga metrik tersebut, beserta nilai t-statistic, p-value, dan signifikansi dari hasil uji.

TABEL V
HASIL UJI STATISTIK PERMODEL MENGGUNAKAN T-TEST

Perbandingan Model	p-value	Signifikansi
PRECISION		
BioBERT vs TinyBERT	0.0694	NO
BioBERT vs ClinicalBERT	0.6891	NO
BioBERT vs Ensemble	0.1594	NO
TinyBERT vs ClinicalBERT	0.1770	NO
TinyBERT vs Ensemble	0.0872	NO
ClinicalBERT vs Ensemble	0.0525	NO
RECALL		
BioBERT vs TinyBERT	0.0495	**YES**
BioBERT vs ClinicalBERT	0.8378	NO
BioBERT vs Ensemble	0.2273	NO
TinyBERT vs ClinicalBERT	0.0849	NO
TinyBERT vs Ensemble	0.0345	**YES**
ClinicalBERT vs Ensemble	0.7285	NO
F1-SCORE		
BioBERT vs TinyBERT	0.0326	**YES**
BioBERT vs ClinicalBERT	0.8826	NO
BioBERT vs Ensemble	0.1755	NO
TinyBERT vs ClinicalBERT	0.1071	NO
TinyBERT vs Ensemble	0.0485	**YES**
ClinicalBERT vs Ensemble	0.2076	NO

Berdasarkan TABEL V, perbandingan antara model-model yang diuji menggunakan T-Test untuk masing-masing metrik evaluasi (*Precision, Recall,* dan *F1-Score*):

1) Precision

Pada metrik Precision, tidak ada perbandingan yang menunjukkan signifikansi yang signifikan. Semua nilai p-value untuk perbandingan antara model *BioBERT, TinyBERT, ClinicalBERT,* dan *Ensemble* lebih besar dari 0.05. Hal ini menunjukkan bahwa perbedaan Precision antara model-model tersebut tidak cukup signifikan untuk dikatakan berbeda secara statistik.

2) Recall

Pada metrik Recall, terdapat dua perbandingan yang menunjukkan signifikansi. Hasil uji T-Test antara *BioBERT* vs *TinyBERT* (p-value = 0.0495) dan *TinyBERT* vs *Ensemble* (p-value = 0.0345) menunjukkan perbedaan yang signifikan. Artinya, ada perbedaan yang cukup besar antara model-model tersebut dalam hal kemampuan untuk mengenali entitas yang relevan, terutama pada Recall.

3) F1-Score

Untuk metrik F1-Score, terdapat dua perbandingan yang juga menunjukkan signifikansi. Perbandingan antara BioBERT vs *TinyBERT* (p-value = 0.0326) dan *TinyBERT* vs *Ensemble* (p-value = 0.0485) menunjukkan bahwa ada perbedaan yang signifikan antara model-model ini dalam hal keseimbangan antara Precision dan Recall.

Untuk menguji signifikansi peningkatan performa antara model-model yang diuji, di penelitian ini menggunakan paired t-test pada metrik *Precision, Recall*, dan *F1-Score*. Hasil uji menunjukkan bahwa meskipun perbedaan pada *Precision* tidak signifikan (p-value > 0.05), terdapat perbedaan signifikan pada *Recall* dan *F1-Score* (p-value = 0.0345 dan p-value = 0.0485), yang mengindikasikan bahwa teknik ensemble voting memberikan peningkatan yang signifikan pada kedua metrik tersebut. Secara khusus, model *Ensemble* menunjukkan kinerja yang lebih unggul dalam beberapa metrik, terutama *Recall* dan *F1-Score*, dibandingkan dengan model *TinyBERT*. Sebaliknya, tidak ditemukan perbedaan signifikan dalam hal *Precision*, yang menunjukkan bahwa meskipun model *Ensemble* mungkin lebih baik dalam beberapa aspek, perbedaan pada *Precision* di antara model-model tersebut relatif kecil.

G. Diskusi tentang Implikasi dan Penerapan

Hasil penelitian ini menunjukkan bahwa pendekatan *ensemble voting* dengan kombinasi BERT-based models memberikan peningkatan signifikan dalam tugas ekstraksi entitas medis, terutama dalam menangani entitas yang jarang muncul. Dengan menggabungkan *BioBERT*, *TinyBERT*, dan *ClinicalBERT*, model ensemble mampu mengatasi kelemahan dari masing-masing model dan memberikan hasil yang lebih stabil. Hasil ini dapat diterapkan dalam aplikasi seperti ekstraksi informasi medis dari rekam medis atau artikel ilmiah, yang dapat mempercepat proses pencarian informasi dan mendukung pengembangan teknologi medis.

H. Kelebihan dan Kekurangan Pendekatan

Kelebihan dari pendekatan *ensemble voting* adalah peningkatan akurasi dan stabilitas yang lebih baik dibandingkan model tunggal, terutama dalam menangani entitas yang lebih jarang. Namun, pendekatan ini membutuhkan lebih banyak sumber daya komputasi dan waktu inferensi yang lebih tinggi dibandingkan dengan model

tunggal. Hal ini perlu dipertimbangkan ketika menerapkan teknik ini pada sistem dunia nyata.

IV. KESIMPULAN

Penelitian ini berhasil meningkatkan kinerja Named Entity Recognition (NER) dalam teks medis dengan menggunakan pendekatan ensemble voting, yang menggabungkan tiga model berbasis BERT: *BioBERT*, *TinyBERT*, dan *ClinicalBERT* untuk meningkatkan akurasi ekstraksi entitas medis pada dataset *BC5CDR*. Hasil evaluasi menunjukkan bahwa Ensemble Voting memberikan performa terbaik dengan *precision*, *recall*, dan *F1-score* yang lebih tinggi dibandingkan dengan model tunggal, terutama dalam mengatasi entitas medis yang lebih jarang. Pendekatan ini terbukti efektif dalam meningkatkan kinerja pengenalan entitas medis dan memberikan potensi aplikasi yang luas di bidang kesehatan, seperti dalam sistem pencarian literatur medis yang lebih efisien, serta sebagai bagian dari clinical decision support untuk mendukung pengambilan keputusan klinis berbasis data.

Namun, meskipun memberikan hasil yang lebih baik, teknik *ensemble* membutuhkan sumber daya komputasi yang lebih besar, sehingga perlu diperhatikan dalam implementasi pada sistem dengan keterbatasan. Penelitian ini membuka peluang untuk pengembangan lebih lanjut, termasuk penerapan teknik data augmentation untuk mengatasi ketidakseimbangan data dan penggunaan model lebih ringan untuk efisiensi waktu komputasi.

Berdasarkan hasil uji T-Test pada metrik *Precision, Recall,* dan *F1-Score*, dapat disimpulkan bahwa meskipun tidak ada perbedaan signifikan antara model-model yang diuji dalam hal *Precision*, terdapat perbedaan signifikan pada *Recall* dan *F1-Score*. Teknik *Ensemble Voting*, yang menggabungkan *BioBERT, TinyBERT,* dan *ClinicalBERT*, memberikan peningkatan signifikan dalam kinerja *Recall* dan *F1-Score*, menunjukkan bahwa model ini lebih baik dalam mengenali entitas medis dan mencapai keseimbangan yang lebih baik antara *precision* dan *recall*. Perbandingan antara *TinyBERT* vs *Ensemble* pada *F1-Score* menunjukkan signifikansi (p-value = 0.0485), yang menegaskan keunggulan *Ensemble* dalam hal ini. Meskipun teknik *Ensemble* memberikan hasil yang lebih baik, perlu diperhatikan bahwa tidak ada perbedaan signifikan pada *Precision* antara model-model tersebut, dan penggunaan teknik *Ensemble* juga memerlukan sumber daya komputasi yang lebih besar dibandingkan dengan model tunggal, yang harus dipertimbangkan dalam implementasinya.

DAFTAR PUSTAKA

[1] M. Triartama Manurung, G. Ngurah, L. Wijayakusuma, I. Putu, and W. Gautama, "Named Entity Recognition for Medical Records of Heart Failure Using a Pre-trained BERT Model," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[2] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[3] P. Su and K. Vijay-Shanker, "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-04642-w.

[4] X. Jiao *et al.*, "TinyBERT: Distilling BERT for Natural Language Understanding," Oct. 2020, [Online]. Available: http://arxiv.org/abs/1909.10351

[5] X. Jiao *et al.*, "Findings of the Association for Computational Linguistics TinyBERT: Distilling BERT for Natural Language Understanding," Nov. 2020.

[6] H. Yu *et al.*, "An intent classification method for questions in 'Treatise on Febrile diseases' based on TinyBERT-CNN fusion model," *Comput Biol Med*, vol. 162, Aug. 2023, doi: 10.1016/j.compbiomed.2023.107075.

[7] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," Nov. 2020, [Online]. Available: http://arxiv.org/abs/1904.05342

[8] B. Yan and M. Pei, "Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation," 2022. [Online]. Available: www.aaai.org

[9] B. Yan and M. Pei, "Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation," 2022. [Online]. Available: www.aaai.org

[10] Y. Jia, H. Wang, Z. Yuan, L. Zhu, and Z. L. Xiang, "Biomedical relation extraction method based on ensemble learning and attention mechanism," *BMC Bioinformatics*, vol. 25, no. 1, p. 333, Dec. 2024, doi: 10.1186/s12859-024-05951-y.

[11] A. M. Bamhdi, I. Abrar, and F. Masoodi, "An ensemble based approach for effective intrusion detection using majority voting," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 2, pp. 664–671, Apr. 2021, doi: 10.12928/TELKOMNIKA.v19i2.18325.

[12] M. A. Naji, S. El Filali, M. Bouhlal, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 481–486. doi: 10.1016/j.procs.2021.07.061.

[13] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Apr. 2021, doi: 10.1016/j.engappai.2022.105151.

[14] S. Masoumi, H. Amirkhani, N. Sadeghian, and S. Shahraz, "Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research," *Syst Rev*, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13643-024-02470-y.

[15] Z. Li *et al.*, "Ensemble pretrained language models to extract biomedical knowledge from literature," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1904–1911, Sep. 2024, doi: 10.1093/jamia/ocae061.

[16] T. Meenachisundaram and M. Dhanabalachandran, "Biomedical Named Entity Recognition Using the SVM Methodologies and bio Tagging Schemes," 2021, doi: 10.37358/Rev.Chim.1949.

[17] Q. Zhang, Y. Sun, L. Zhang, Y. Jiao, and Y. Tian, "Named entity recognition method in health preserving field based on BERT," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 212–220. doi: 10.1016/j.procs.2021.03.010.

[18] A. C. Mazari, N. Boudoukhani, and A. Djeffal, "BERT-based ensemble learning for multi-aspect hate speech detection," *Cluster Comput*, vol. 27, no. 1, pp. 325–339, Feb. 2024, doi: 10.1007/s10586-022-03956-x.

[19] C. D. Nafanda and A. Salam, "Optimalisasi Model BioBERT untuk Pengenalan Entitas pada Teks Medis dengan Conditional Random Fields (CRF)," *Technology and Science (BITS)*, vol. 6, no. 4, 2025, doi: 10.47065/bits.v6i4.7042.

[20] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," *Jurnal Teknik*, vol. 12, Oct. 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index