CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGGUNAKAN *PRE-TRAINED WORD EMBEDDINGS* DAN *DEEP LEARNING*

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH :

MUHAMMAD AZRIEL APRIADI

09011282126078

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2025

HALAMAN PENGESAHAN

SKRIPSI

CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGGUNAKAN *PRE-TRAINED WORD EMBEDDINGS* DAN *DEEP LEARNING*

Sebagai salah satu syarat untuk penyelesaian studi di

Program Studi S1 Sistem Komputer

Oleh:

Muhammad Azriel Apriadi

09011282126078

Pembimbing 1 : **Dr. Firdaus, ST., M.Kom.**
NIP. 197801212008121003

Mengetahui

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T.
196612032006041001

AUTHENTICATION PAGE

SKRIPSI

CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA USING PRE-TRAINED WORD EMBEDDINGS AND DEEP LEARNING

As one of the requirements for completing the

Bachelor's Degree Program in Computer Systems

By:

Muhammad Azriel Apriadi

09011282126078

Supervisor 1 : **Dr. Firdaus, ST., M.Kom.**
NIP. 197801212008121003

Approved by,

Head of Computer System Department

Dr. Ir. Sukemi, M.T.
196612032006041001

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 23 Mei 2025

Tim Penguji :

1. Ketua : Prof. Dr. Erwin, S.Si., M.Si.
2. Penguji : Prof. Dr. Ir. Bambang Tutuko, M.T.
3. Pembimbing : Dr. Firdaus, S.T., M.Kom.

Mengetahui, 17/6/25

Ketua Jurusan Sistem Komputer

KEMENTERIAN PENDIDIKAN TINGGI SAINS DAN TEKNOLOGI
UNIVERSITAS SRIWIJAYA
JURUSAN SISTEM KOMPUTER

Dr. Ir. Sukemi, M.T.

NIP. 196612032006041001

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini :

Nama : Muhammad Azriel Apriadi

NIM : 09011282126078

Judul : *Clinical Named Entity Recognition* pada Data Biomedis Menggunakan *Pre-trained Word Embeddings* dan *Deep Learning*

Hasil Pengecekan *Software* Turnitin: 2%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian, pernyataan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.

Indralaya, 20 Juni 2025

MUHAMMAD AZRIEL APRIADI
NIM. 09011282126078

HALAMAN PERSEMBAHAN

1. Untuk Ayah dan Ibu, dengan rasa hormat dan cinta yang tak terhingga, saya persembahkan karya ini. Terima kasih atas setiap doa yang menguatkan, setiap dukungan yang setia mengiringi, dan setiap pengorbanan yang tak pernah kalian ungkapkan. Tanpa kalian, perjalanan ini tak akan pernah sampai di titik ini. Semoga pencapaian ini menjadi wujud kecil dari rasa terima kasih saya atas segalanya yang telah kalian berikan, dan menjadi kebahagiaan yang turut kalian rasakan.
2. Untuk Kakakku dan Adikku, terima kasih atas kebersamaan, dukungan, dan semangat yang selalu kalian berikan. Perjalanan ini lebih ringan karena ada kalian yang setia menemani, dalam suka maupun duka. Kehadiran kalian menjadi penguat di setiap langkah, dan sumber keceriaan di tengah lelah. Skripsi ini aku persembahkan juga untuk kalian, bagian penting dalam cerita ini.

MOTTO

"Sometimes even to live is an act of courage."

(Seneca)

"Only the educated are free."

(Epictetus)

"Tidak semua keberanian tampak seperti kemenangan besar. Kadang, keberanian sejati adalah terus hidup saat segalanya terasa sunyi, berjalan pelan saat dunia berlari. Kita tak harus menjadi luar biasa untuk bernilai cukup terus belajar, tumbuh, dan mencoba. Dalam proses itulah kita menemukan bukan hanya ilmu, tetapi juga kebebasan: untuk memahami diri, memilih arah, dan bangkit dari luka yang tak terlihat. Sebab pada akhirnya, keteguhan berpijak di tengah keterbatasan adalah bentuk kemerdekaan yang paling utuh."

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh,

Segala puji dan syukur kita panjatkan ke hadirat Allah SWT yang telah melimpahkan rahmat, kasih sayang, dan karunia-Nya, sehingga penulis dapat menyelesaikan tugas akhir dengan judul "*Clinical Named Entity Recognition* pada Data Biomedis Menggunakan *Pre-trained Word Embeddings* dan *Deep Learning*".

Selama proses pembuatan dan penulisan tugas akhir ini, penulis telah menerima banyak bantuan dan dukungan dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Orang tua, saudara, dan seluruh keluarga besar yang telah mendoakan, memberikan motivasi, dan mendukung penulis.
2. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Bapak Dr. Ir. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Universitas Sriwijaya.
4. Bapak Dr. Firdaus, S.T., M.Kom. sebagai Dosen Pembimbing Tugas Akhir yang telah meluangkan waktu untuk membimbing, memberikan saran, motivasi, dan bimbingan terbaik kepada penulis dalam menyelesaikan Tugas Akhir ini.
5. Ibu Prof. Dr. Ir. Siti Nurmaini, M.T., Ph.D. selaku *Head of Intelligent System Research Group* (ISysRG), yang telah memberikan kesempatan berharga untuk bergabung dan menjadi bagian dari *research group* ini.
6. Ibu Anggun Islami, M.Kom., Ibu Dr. Ade Iriani Sapitri, M.Kom., Ibu Annisa Darmawahyuni, S.Kom., Ibu Akhiar Wista Arum, M.Kom., M.T., dan Bapak Naufal Rachmatullah sebagai mentor di ISysRG.
7. Bapak Sutarno, S.T, M.T, selaku dosen pembimbing akademik.
8. Staff administrasi Fakultas Ilmu Komputer Universitas Sriwijaya Jurusan Sistem Komputer Kampus Indralaya yang telah memberikan kemudahan

dalam hal administrasi sehingga penulis dapat membuat proposal tugas akhir ini dengan lancar.

9. Teman-teman satu divisi *text processing*, Keisyah Sabinatullah Qur'aini, Indah Gala Putri, Tria Lailani, Tiara Oktarina, dan seluruh anggota IsysRG lainnya yang telah menemani dimasa skripsi
10. Grup "Botanisme" Ade, Adam, Arif, Fakhrul, Farhan, Quddus, dan Reihan, terima kasih sudah menjadi tempat bercerita, bertukar informasi, dan saling mendukung sepanjang perjalanan ini
11. Teman-teman seperjuangan Jurusan Sistem Komputer Angkatan 2021, terutama teman-teman kelas SKB, dan Muhammad Rafi Rizqullah, yang telah menjadi tempat saling berbagi ilmu dan banyak membantu dalam berbagai hal selama perkuliahan.
12. Seluruh pihak yang membantu penulis selama proses pembuatan tugas akhir yang tidak dapat disebutkan satu persatu.
13. Almamater

Penulis menyadari bahwa laporan tugas akhir ini masih jauh dari sempurna. Oleh karena itu, penulis menyadari adanya banyak kekurangan dan kesalahan dalam penulisan laporan ini. Penulis sangat mengharapkan kritik dan saran yang membangun untuk perbaikan laporan-laporan di masa yang akan datang. Semoga laporan ini dapat bermanfaat bagi semua orang.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, 20 Juni 2025

Penulis,

Muhammad Azriel Apriadi
NIM. 09011282126078

CLINICAL NAMED ENTITY RECOGNITION PADA DATA BIOMEDIS MENGGUNAKAN *PRE-TRAINED WORD EMBEDDINGS* DAN *DEEP LEARNING*

MUHAMMAD AZRIEL APRIADI (09011282126078)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email: aapriadiazriel@gmail.com

ABSTRAK

Pertumbuhan pesat data biomedis digital memunculkan tantangan dalam pengelolaan dan ekstraksi informasi dari teks medis tidak terstruktur. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model *Clinical Named Entity Recognition* (CNER) dengan mengombinasikan *pre-trained word embeddings* dan algoritma *deep learning*. Tiga dataset biomedis digunakan, yaitu JNLPBA, NCBI-*Disease*, dan BC2GM. Eksperimen dilakukan dalam dua tahap: tahap pertama membandingkan performa kombinasi GloVe-BiLSTM, ELMo-BiLSTM, dan BERT-BiLSTM; tahap kedua mengevaluasi BERT-BiLSTM dan PubMed2MBERT-BiLSTM dengan pendekatan *fine-tuning* dan strategi *early stopping*. Evaluasi menggunakan *macro average* F1-*Score* menunjukkan bahwa *contextual embeddings* secara konsisten mengungguli *static embeddings*, dengan GloVe mencatat performa terendah. Model berbasis *transformer* seperti BERT dan PubMed2MBERT melampaui ELMo berkat kemampuan *self-attention* dalam menangkap relasi antar-token. PubMed2MBERT-BiLSTM, yang dilatih khusus untuk domain biomedis, menunjukkan performa terbaik di seluruh dataset, menegaskan efektivitas model *domain-specific* dalam pengenalan entitas medis.

Kata Kunci : *Clinical Named Entity Recognition*, *Deep Learning*, *Pre-trained Word Embeddings*, Teks Biomedis, GloVe, ELMo, BERT, PubMed2MBERT, *Transformer,* BiLSTM.

CLINICAL NAMED ENTITY RECOGNITION ON BIOMEDICAL DATA USING PRE-TRAINED WORD EMBEDDINGS AND DEEP LEARNING

MUHAMMAD AZRIEL APRIADI (09011282126078)

Computer System Department, Computer Science Faculty, Sriwijaya University

Email : aapriadiazriel@gmail.com

ABSTRACT

The rapid growth of digital biomedical data has posed significant challenges in managing and extracting information from unstructured medical texts. This study aims to develop and evaluate a Clinical Named Entity Recognition (CNER) model by combining pre-trained word embeddings with deep learning architectures. Three biomedical datasets were used: JNLPBA, NCBI-Disease, and BC2GM. The experiments were conducted in two stages: the first stage compared the performance of GloVe-BiLSTM, ELMo-BiLSTM, and BERT-BiLSTM combinations; the second stage evaluated BERT-BiLSTM and PubMed2MBERT-BiLSTM models using fine-tuning and early stopping strategies. Evaluation using macro average precision, recall, and F1-Score shows that contextual embeddings consistently outperform static embeddings, with GloVe yielding the lowest performance. Transformer-based models like BERT and PubMed2MBERT outperform ELMo due to their self-attention mechanism that better captures token relationships. PubMed2MBERT-BiLSTM, pretrained in the biomedical domain, achieved the best performance across all datasets, highlighting the effectiveness of domain-specific models in medical entity recognition.

Keywords : *Clinical Named Entity Recognition, Deep Learning, Pre-trained Word Embeddings, Biomedical Text, GloVe, ELMo, BERT, PubMed2MBERT, Transformer, BiLSTM.*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1. Latar Belakang

Dalam beberapa tahun terakhir, pertumbuhan dokumen dan data di bidang biomedis berkembang sangat pesat [1]. Perkembangan ini didorong oleh perubahan sistem pencatatan dari kertas ke format digital, yang memudahkan dokumentasi informasi medis dan menyebabkan peningkatan jumlah data secara signifikan [2][3]. Dikarenakan banyaknya data yang ada muncul tantangan tersendiri dalam hal pengolahan dan analisis informasi yang tersedia. Para staf klinis menghabiskan banyak waktu untuk mengelola dan menganalisis dokumen yang melimpah, yang tidak hanya berkontribusi terhadap kelelahan mereka dan menyita lebih banyak waktu, tetapi juga meningkatkan risiko kesalahan akibat analisis manual [3][4].

Berkembangnya teknologi *artificial intelligence* dan *machine learning* terutama dibidang *Natural Language Processing* (NLP) memberikan solusi terkait permasalahan ekstraksi informasi, dalam kasus ini terhadap teks biomedis [3][4]. Salah satu metode yang dapat dilakukan adalah *Named Entity Recogniton* (NER). NER berfungsi untuk mengidentifikasi nama-nama spesifik dalam teks yang sesuai dengan kategori tertentu, sehingga memungkinkan kita untuk secara efektif mengekstraksi data yang relevan dan mengelompokkan entitas tersebut ke dalam berbagai kelas [1][5][6]. Dalam konteks biomedis, NER dapat digunakan untuk mengidentifikasi dan mengekstrak informasi penting seperti diagnosis penyakit, nama obat, prosedur medis, serta faktor risiko yang dapat mempengaruhi kesehatan pasien, jenis NER ini biasa disebut *Clinical Named Entity Recognition* atau CNER [1][3]. Dengan demikian, NER berperan penting dalam analisis data medis, mendukung pengambilan keputusan yang lebih baik dan pemahaman yang lebih mendalam tentang informasi kesehatan.

Membuat sistem NER dapat dilakukan dengan berbagai pendekatan, seperti *rule-based systems, unsupervised learning, feature-based supervised learning,* dan *deep learning* [5]. Saat ini, pendekatan *deep learning* menjadi tren dominan karena kemampuannya untuk belajar representasi kompleks secara otomatis,

menangkap ketergantungan konteks yang lebih dalam, dan dilatih dalam satu langkah, sehingga meningkatkan akurasi dan efisiensi sistem NER [2][5]. Selanjutnya, beberapa studi menunjukkan bahwa penggunaan *pre-trained word embeddings* seperti Word2Vec dan GloVe dapat meningkatkan performa dalam berbagai tugas NLP, termasuk klasifikasi teks, analisis sentimen, dan NER, dengan hasil evaluasi yang seringkali menunjukkan akurasi lebih tinggi dibandingkan model tanpa *pre-trained word embeddings* [5][6].

Penelitian ini bertujuan untuk mengimplementasikan model CNER pada data biomedis dengan menggunakan metode *deep learning* dan *pre-trained word embedding*. Pendekatan ini diharapkan dapat meningkatkan akurasi dalam pengenalan entitas medis, sehingga memungkinkan ekstraksi informasi yang lebih tepat dan efisien dari teks rekam medis yang tidak terstruktur. Dengan mengintegrasikan teknologi *deep learning* dan *pre-trained word embedding*, model ini diharapkan mampu menangkap relasi dan konteks yang lebih kompleks, yang sangat penting untuk aplikasi dalam riset klinis dan manajemen informasi kesehatan, serta berkontribusi terhadap deteksi dini penyakit dan peningkatan kualitas pelayanan kesehatan.

1.2. Rumusan Masalah

Rumusan masalah dalam penelitian ini berfungsi untuk mengidentifikasi isu-isu utama yang akan diselesaikan serta memberikan panduan dalam pelaksanaan penelitian. Adapun perumusan masalah yang diajukan adalah sebagai berikut:

1. Bagaimana proses penerapan model CNER dengan kombinasi *pre-trained word embedding*s dan algoritma *deep learning*?
2. Bagaimana pengaruh variasi *pre-trained word embedding*s dan arsitektur *deep learning* terhadap hasil CNER?
3. Bagaimana performa *static* dan *contextual word embeddings saat* dikombinasikan dengan model *deep learning* untuk tugas CNER?
4. Bagaimana performa *domain-specific word embeddings* saat dikombinasikan dengan model *deep learning* untuk tugas CNER?

1.3. Batasan Masalah

Batasan masalah bertujuan untuk memperjelas ruang lingkup penelitian ini agar tetap terfokus dan efektif, serta menghindari pembahasan yang terlalu luas. Dengan menetapkan batasan ini, penelitian diharapkan dapat berjalan lebih sistematis dan sesuai dengan tujuan yang telah ditentukan. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya mencakup pembuatan model untuk CNER menggunakan *pre-trained word embeddings* dan algoritma *deep learning.* Model atau pendekatan lain yang tidak berbasis *deep learning* dan *pre-trained word embeddings* tidak akan dibahas dalam penelitian ini.
2. Dataset yang digunakan hanya mencakup data berbentuk teks dengan anotasi yang didapat dari sumber terbuka. Data dalam format lain seperti gambar, audio, atau dalam bahasa selain bahasa Inggris, tidak termasuk dalam lingkup penelitian ini.
3. Entitas yang dapat dikenali oleh model yang dibuat pada penelitian ini hanya dapat mengenali entitis medis tertentu, seperti penyakit, gejala, obat-obatan, sel dan genetik. Entitas umum yang biasa ada pada NER seperti nama orang, organisasi, lokasi, dan lain-lain tidak akan dipertimbangkan dalam penelitian ini.
4. Penelitian ini membatasi evaluasi kinerja model CNER pada tiga metrik utama, yakni F1-*Score*, *precision*, dan *recall*, karena ketiganya dianggap lebih efektif dalam mengukur performa CNER, mampu menyeimbangkan kesalahan prediksi, serta memberikan hasil evaluasi yang lebih representatif dibandingkan metrik *accuracy*.

1.4. Tujuan

Adapun tujuan dari penelitian ini yaitu:

1. Menerapkan model CNER untuk data biomedis dengan mengombinasikan *pre-trained word embeddings* dan model *deep learning*.
2. Menganalisis pengaruh variasi *pre-trained word embeddings* dan arsitektur *deep learning* terhadap performa model CNER.

3. Membandingkan performa *static* dan *contextual word embeddings* untuk tugas CNER saat dikombinasikan dengan model *deep learning*.
4. Mengevaluasi performa *domain-specific word embeddings* untuk tugas CNER saat dikombinasikan dengan model *deep learning*.

1.5. Metodologi Penelitian

Pada penelitian ini, metodologi yang diterapkan mencakup berbagai pendekatan yang dirancang untuk memastikan bahwa proses penelitian berjalan secara terstruktur dan menghasilkan temuan yang valid dan dapat diandalkan. Metodologi ini terdiri dari lima tahapan utama diantaranya :

1.5.1. Metode Studi Pustaka dan Literatur

Metode ini dilakukan dengan mencari dan mengumpulkan referensi dari literatur, khususnya jurnal-jurnal terpercaya, yang berkaitan dengan "*Clinical Named Entity Recognition* pada Data Biomedis Menggunakan *Pre-trained word embedding*s dan *Deep Learning*".

1.5.2. Metode Konsultasi

Metode ini melibatkan konsultasi dengan pihak-pihak yang memiliki pengetahuan serta wawasan yang baik dalam mengatasi permasalahan yang ditemui pada penulisan tugas akhir "*Clinical Named Entity Recognition* pada Data Biomedis Menggunakan *Pre-trained word embedding*s dan *Deep Learning*".

1.5.3. Metode Pembuatan Model

Metode ini melibatkan proses pengumpulan dataset, pengolahan dataset, memiliih *pre-trained word embedding* dan algoritma *deep learning*, dan melakukan *training* model menggunakan dataset dan kombinasi model.

1.5.4. Metode Pengujian

Metode pengujian ini bertujuan untuk mengevaluasi model CNER yang telah dibuat dengan menggunakan berbagai metrik evaluasi untuk memastikan apakah performa model tersebut baik.

1.5.5. Metode Analisa dan Kesimpulan

Hasil pengujian dalam tugas akhir ini akan dievaluasi kelemahannya. Dengan memahami kelemahan tersebut, kita dapat mengembangkan rekomendasi untuk perbaikan dan peningkatan model di masa depan, sehingga penelitian selanjutnya dapat memanfaatkan temuan ini untuk mencapai hasil yang lebih optimal dan relevan dalam aplikasi dunia nyata.

1.6. Sistematika Penulisan

Laporan ini menggunakan sistematika penulisan sebagai berikut :

BAB I PENDAHULUAN

Bagian ini akan menjelaskan alasan dilakukannya penelitian serta apa yang ingin dicapai dan batasan-batasan yang diterapkan dalam penelitian. Selain itu, bab ini juga mencakup metode penulisan yang digunakan dan sistematika penulisan laporan secara keseluruhan, untuk memberikan gambaran menyeluruh mengenai struktur penelitian yang dilakukan.

BAB II TINJAUAN PUSTAKA

Bab ini mengulas penelitian-penelitian terdahulu yang berhubungan dengan topik penelitian sebagai dasar teoretis dan referensi untuk mengembangkan penelitian lebih lanjut. Selain itu, bab ini juga mencakup konsep-konsep dasar yang digunakan dalam penelitian, seperti NER, *pre-trained word embeddings*, dan algoritma *deep learning*.

BAB III METODOLOGI PENELITIAN

Bab ini menguraikan metodologi penelitian yang digunakan untuk mencapai tujuan penelitian, mencakup langkah-langkah yang ditempuh dalam pengumpulan data, pengolahan data, pembuatan model, serta metode evaluasi yang diterapkan. Metode yang digunakan dirancang untuk memastikan bahwa penelitian dilakukan secara sistematis dan dapat dipertanggungjawabkan.

BAB IV HASIL DAN ANALISIS

Bab ini berisi tentang hasil dari penelitian yang telah dilakukan serta analisis terhadap hasil tersebut. Hasil penelitian mencakup performa model yang telah dibangun berdasarkan berbagai metrik evaluasi. Selain itu, bab ini juga membahas interpretasi dari hasil yang diperoleh serta faktor-faktor yang mempengaruhi performa model.

BAB V KESIMPULAN

Bab ini menyajikan kesimpulan dari hasil penelitian yang telah dilakukan serta rekomendasi untuk penelitian selanjutnya. Kesimpulan ini mencakup jawaban atas rumusan masalah yang telah disusun dan pencapaian tujuan penelitian. Selain itu, bab ini juga menyampaikan saran yang dapat diterapkan pada penelitian lanjutan untuk meningkatkan kualitas model dan hasil yang diperoleh.

DAFTAR PUSTAKA

Bagian ini memuat daftar referensi yang digunakan dalam penelitian ini, yang terdiri dari buku, jurnal, artikel, dan sumber-sumber lain yang relevan dan kredibel.

LAMPIRAN

Bagian lampiran dalam laporan ini memuat dokumentasi yang relevan dengan penelitian, termasuk foto, tabel, dan catatan tambahan. Dokumentasi ini bertujuan untuk memberikan informasi lebih mendalam serta bukti pendukung yang dapat memperkuat hasil dan analisis dalam laporan utama.

DAFTAR PUSTAKA

[1] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decision Analytics Journal*, vol. 10, 2024, doi: 10.1016/j.dajour.2024.100426.

[2] G. Hou, Y. Jian, Q. Zhao, X. Quan, and H. Zhang, "Language model based on deep learning network for biomedical named entity recognition," *Methods*, vol. 226, pp. 71–77, Jun. 2024, doi: 10.1016/j.ymeth.2024.04.013.

[3] D. Fraile Navarro *et al.*, "Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review," 2023. doi: 10.1016/j.ijmedinf.2023.105122.

[4] J. Ravikumar and P. Ramakanth Kumar, "Machine learning model for clinical named entity recognition," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, 2021, doi: 10.11591/ijece.v11i2.pp1689-1696.

[5] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans Knowl Data Eng*, vol. 34, no. 1, 2022, doi: 10.1109/TKDE.2020.2981314.

[6] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artif Intell Rev*, vol. 56, no. 9, 2023, doi: 10.1007/s10462-023-10419-1.

[7] S. Srivastava, B. Paul, and D. Gupta, "Study of Word Embeddings for Enhanced Cyber Security Named Entity Recognition," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 449–460. doi: 10.1016/j.procs.2023.01.027.

[8] R. E. Ramos-Vargas, I. Román-Godínez, and S. Torres-Ramos, "Comparing general and specialized word embeddings for biomedical named entity recognition," *PeerJ Comput Sci*, vol. 7, 2021, doi: 10.7717/peerj-cs.384.

[9] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, "BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-05051-9.

[10] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, 2023, doi: 10.1007/s11042-022-13428-4.

[11] S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and S. Fouad, “A review of natural language processing in contact centre automation,” *Pattern Analysis and Applications*, vol. 26, no. 3, 2023, doi: 10.1007/s10044-023-01182-8.

[12] T. X. Sun, X. Y. Liu, X. P. Qiu, and X. J. Huang, “Paradigm Shift in Natural Language Processing,” 2022. doi: 10.1007/s11633-022-1331-6.

[13] D. H. Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. M. Sadeeq, and K. H. Sharif, “A State of Art for Semantic Analysis of Natural Language Processing,” *Qubahan Academic Journal*, vol. 1, no. 2, 2021, doi: 10.48161/qaj.v1n2a44.

[14] A. Galassi, M. Lippi, and P. Torroni, “Attention in Natural Language Processing,” *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 10, 2021, doi: 10.1109/TNNLS.2020.3019893.

[15] P. Liu, Y. Guo, F. Wang, and G. Li, “Chinese named entity recognition: The state of the art,” *Neurocomputing*, vol. 473, pp. 37–53, Feb. 2022, doi: 10.1016/j.neucom.2021.10.101.

[16] Z. Nasar, S. W. Jaffry, and M. K. Malik, “Named Entity Recognition and Relation Extraction: State-of-The-Art,” *ACM Comput Surv*, vol. 54, no. 1, Apr. 2021, doi: 10.1145/3445965.

[17] Q. H. Ngo, T. Kechadi, and N. A. Le-Khac, “Domain specific entity recognition with semantic-based deep learning approach,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3128178.

[18] Warto, Muljono, Purwanto, and E. Noersasongko, “Improving Named Entity Recognition in Bahasa Indonesia with Transformer-Word2Vec-CNN-Attention Model,” *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 4, 2023, doi: 10.22266/ijies2023.0831.53.

[19] A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, and N. Dragoni, “BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling,” *Applied Sciences (Switzerland)*, vol. 12, no. 3, 2022, doi: 10.3390/app12030976.

[20] S. H. Jeon and S. Cho, “Edge Weight Updating Neural Network for Named Entity Normalization,” *Neural Process Lett*, vol. 55, no. 5, 2023, doi: 10.1007/s11063-022-11102-2.

[21] Y. Tian, W. Shen, Y. Song, F. Xia, M. He, and K. Li, “Improving biomedical named entity recognition with syntactic information,” *BMC Bioinformatics*, vol. 21, no. 1, 2020, doi: 10.1186/s12859-020-03834-6.

[22] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, “A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning,” *Technologies (Basel)*, vol. 10, no. 3, 2022, doi: 10.3390/technologies10030057.

[23] M. A. Palomino and F. Aider, “Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis,” *Applied Sciences (Switzerland)*, vol. 12, no. 17, 2022, doi: 10.3390/app12178765.

[24] S. Nazir, M. Asif, M. Rehman, and S. Ahmad, “Machine learning based framework for fine-grained word segmentation and enhanced text normalization for low resourced language,” *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/peerj-cs.1704.

[25] R. Wolert and M. Rawski, “Email Phishing Detection with BLSTM and Word Embeddings,” *International Journal of Electronics and Telecommunications*, vol. 69, no. 3, 2023, doi: 10.24425/ijet.2023.146496.

[26] Nadir Hussain, Dr. Sheikh Muhammad Saqib, Hamza Arif, and Muhammad Usman Gurmani, “Detection of Questions from Text Data Using LSTM-Deep Learning Model,” *VAWKUM Transactions on Computer Sciences*, vol. 12, no. 1, 2024, doi: 10.21015/vtcs.v12i1.1655.

[27] A. Ahmed and M. A. Yousuf, “Sentiment analysis on bangla text using long short-term memory (lstm) recurrent neural network,” in *Advances in Intelligent Systems and Computing*, 2021. doi: 10.1007/978-981-33-4673-4_16.

[28] V. R. Joseph, “Optimal ratio for data splitting,” *Stat Anal Data Min*, vol. 15, no. 4, 2022, doi: 10.1002/sam.11583.

[29] R. Karsi, M. Zaim, and J. El Alami, “Leveraging pre-trained contextualized word embeddings to enhance sentiment classification of drug reviews,” *Revue d’Intelligence Artificielle*, vol. 35, no. 4, 2021, doi: 10.18280/ria.350405.

[30] M. A. H. Wadud, M. F. Mridha, and M. M. Rahman, “Word Embedding Methods for Word Representation in Deep Learning for Natural Language Processing,” *Iraqi Journal of Science*, vol. 63, no. 3, 2022, doi: 10.24996/ijs.2022.63.3.37.

[31] M. A. Haq, M. A. R. Khan, and M. Alshehri, “Insider Threat Detection Based on NLP Word Embedding and Machine Learning,” *Intelligent Automation and Soft Computing*, vol. 33, no. 1, 2022, doi: 10.32604/iasc.2022.021430.

[32] J. M. Imperial, “BERT Embeddings for Automatic Readability Assessment,” in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021. doi: 10.26615/978-954-452-072-4_069.

[33] M. Alawad *et al.*, “Privacy-Preserving Deep Learning NLP Models for Cancer Registries,” *IEEE Trans Emerg Top Comput*, vol. 9, no. 3, 2021, doi: 10.1109/TETC.2020.2983404.

[34] J. Chai and A. Li, “Deep Learning in Natural Language Processing: A State-of-the-Art Survey,” in *Proceedings - International Conference on Machine Learning and Cybernetics*, 2019. doi: 10.1109/ICMLC48188.2019.8949185.

[35] S. K. Hong and J. G. Lee, “DTranNER: Biomedical named entity recognition with deep learning-based label-label transition model,” *BMC Bioinformatics*, vol. 21, no. 1, 2020, doi: 10.1186/s12859-020-3393-1.

[36] R. Adipradana, B. P. Nayoga, R. Suryadi, and D. Suhartono, “Hoax analyzer for indonesian news using rnns with fasttext and glove embeddings,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, 2021, doi: 10.11591/eei.v10i4.2956.

[37] N. Patwardhan, S. Marrone, and C. Sansone, “Transformers in the Real World: A Survey on NLP Applications,” 2023. doi: 10.3390/info14040242.

[38] A. Rahali and M. A. Akhloufi, “End-to-End Transformer-Based Models in Textual-Based NLP,” 2023. doi: 10.3390/ai4010004.

[39] P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, “A survey on recent named entity recognition and relationship extraction techniques on clinical texts,” 2021. doi: 10.3390/app11188319.