ANALISIS PENGARUH *LATENT DIRICHLET ALLOCATION* DAN *GRAPH CONVOLUTIONAL NETWORK* PADA MODEL X UNTUK *NAMED ENTITY RECOGNITION* BERITA PENYAKIT MENULAR BERBAHASA INDONESIA

Skripsi

Oleh

ERIN ELFITRIANI
NPM. 2217031156

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

ABSTRACT

ANALYSIS OF THE EFFECT OF LATENT DIRICHLET ALLOCATION AND GRAPH CONVOLUTIONAL NETWORK ON MODEL X FOR NAMED ENTITY RECOGNITION OF INDONESIAN-LANGUAGE INFECTIOUS DISEASE NEWS

By

Erin Elfitriani

The increasing volume of Indonesian digital text has driven the need for automatic information extraction systems, particularly Named Entity Recognition (NER). Class imbalance and the complexity of entity structures remain challenges that may affect the consistency of entity type recognition. This study aims to analyze the performance of Model X, namely the IndoBERT–BiLSTM hybrid model, and to evaluate the effect of integrating Latent Dirichlet Allocation (LDA) topic features and the structural relations of Graph Convolutional Network (GCN) on NER performance based on entity types. The evaluation was conducted using precision, recall, F1-score, accuracy, and macro-average F1-score metrics. The results show that Model X achieved a macro F1-score of 0.9127 with an accuracy of 0.9681. The integration of LDA improved the recall value but reduced precision, resulting in a macro F1-score of 0.7119. The model integrated with GCN demonstrated a better balance between precision and recall, achieving a macro F1-score of 0.8611. Meanwhile, the combination of LDA and GCN produced high recall across all entity types, but the decline in precision led to a macro F1-score of 0.6709. These findings indicate differences in performance characteristics across the model scenarios, where Model X yielded the highest aggregate performance, while the integration of GCN showed more consistent entity detection capability compared with the other approaches.

Keywords: Named Entity Recognition, IndoBERT, BiLSTM, Latent Dirichlet Allocation, Graph Convolutional Network

ABSTRAK

ANALISIS PENGARUH *LATENT DIRICHLET ALLOCATION* DAN *GRAPH CONVOLUTIONAL NETWORK* PADA MODEL X UNTUK *NAMED ENTITY RECOGNITION* BERITA PENYAKIT MENULAR BERBAHASA INDONESIA

Oleh

Erin Elfitriani

Peningkatan volume teks digital berbahasa Indonesia mendorong kebutuhan akan sistem ekstraksi informasi otomatis, khususnya *Named Entity Recognition* (NER). Ketidakseimbangan kelas dan kompleksitas struktur entitas masih menjadi tantangan yang dapat memengaruhi konsistensi pengenalan tipe entitas. Penelitian ini bertujuan untuk menganalisis kinerja model X, yaitu model hibrida IndoBERT–BiLSTM, serta mengevaluasi pengaruh integrasi fitur topik *Latent Dirichlet Allocation* (LDA) dan relasi struktural *Graph Convolutional Network* (GCN) terhadap performa NER berdasarkan tipe entitas. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, *F1-score*, *accuracy*, dan *macro average F1-score*. Hasil penelitian menunjukkan bahwa model X memperoleh *macro F1-score* sebesar 0,9127 dengan akurasi 0,9681. Integrasi LDA meningkatkan nilai *recall*, namun menurunkan *precision* sehingga *macro F1-score* menjadi 0,7119. Model dengan integrasi GCN menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall* dengan *macro F1-score* sebesar 0,8611. Sementara itu, kombinasi LDA dan GCN menghasilkan *recall* yang tinggi pada seluruh tipe entitas, tetapi penurunan *precision* menyebabkan *macro F1-score* menjadi 0,6709. Hasil penelitian ini menunjukkan adanya perbedaan karakteristik kinerja antar skenario model, di mana model X memberikan performa agregat tertinggi, sedangkan integrasi GCN menunjukkan kemampuan deteksi entitas yang lebih konsisten dibandingkan pendekatan lainnya.

Kata-kata kunci: *Named Entity Recognition*, IndoBERT, BiLSTM, *Latent Dirichlet Allocation*, *Graph Convolutional Network*

ANALISIS PENGARUH *LATENT DIRICHLET ALLOCATION* DAN *GRAPH CONVOLUTIONAL NETWORK* PADA MODEL X UNTUK *NAMED ENTITY RECOGNITION* BERITA PENYAKIT MENULAR BERBAHASA INDONESIA

ERIN ELFITRIANI

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2026

Judul Skripsi	:	**ANALISIS PENGARUH *LATENT DIRICHLET ALLOCATION* DAN *GRAPH CONVOLUTIONAL NETWORK* PADA MODEL X UNTUK *NAMED ENTITY RECOGNITION* BERITA PENYAKIT MENULAR BERBAHASA INDONESIA**
Nama Mahasiswa	:	Erin Elftriani
Nomor Pokok Mahasiswa	:	**2217031156**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Dian Kurniasari, S.Si., M.Sc.	**Dr. Purnomo Husnul Khotimah, M.T.**
NIP. 196903051996032001	NIP. 198003232005022002

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Dian Kurniasari, S.Si., M.Sc.** .....

Sekretaris : **Dr. Purnomo Husnul Khotimah, M.T.** .....

Penguji

Bukan Pembimbing : **Prof. Dra. Wamiliana, MA., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.

NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **14 April 2026**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Erin Elfitriani**
Nomor Pokok Mahasiswa	:	**2217031156**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Analisis Pengaruh *Latent Dirichlet Allocation* dan *Graph Convolutional Network* pada Model X untuk *Named Entity Recognition* Berita Penyakit Menular Berbahasa Indonesia**

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri. Apabila kemudian hari terbukti bahwa skripsi inni merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 14 April 2026

Penulis

Erin Elfitriani

RIWAYAT HIDUP

Penulis bernama lengkap Erin Elfitriani, lahir di Prabumulih pada tanggal 24 November 2003. Penulis merupakan putri bungsu dari almarhum Irhamjaya dan Ibu Rusyati. Penulis menempuh pendidikan dasar di SD Negeri Tanjung Tiga dan lulus pada tahun 2015. Pendidikan menengah pertama ditempuh di SMP Negeri 3 Rebang Tangkas dan lulus pada tahun 2018, kemudian melanjutkan pendidikan menengah atas di SMA Negeri 1 Rebang Tangkas dan lulus pada tahun 2021.

Pada tahun 2022, penulis melanjutkan studi ke jenjang perguruan tinggi pada Program Studi S1 Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung melalui jalur Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN). Selama menjadi mahasiswi, penulis aktif mengikuti kegiatan Unit Kegiatan Mahasiswa Universitas Lampung Sains dan Teknologi (UKM U Saintek Unila) pada bidang Sumber Daya Manusia (SDM).

Selain kegiatan organisasi, penulis juga aktif dalam kegiatan akademik. Penulis pernah diamanahkan sebagai asisten dosen pada mata kuliah Statistika Dasar untuk Program Studi Biologi FMIPA Universitas Lampung. Penulis juga pernah mengikuti lomba esai GEMASTIK (Pagelaran Mahasiswa Nasional Bidang Teknologi Informasi dan Komunikasi). Selain itu, penulis pernah menjadi mentor pada mata kuliah Pengantar Teori Peluang dan Aljabar Linear Elementer bagi mahasiswa baru Program Studi Matematika.

Pada tahun 2025, penulis melaksanakan Praktik Kerja Lapangan (PKL) serta mengikuti program Merdeka Belajar Kampus Merdeka (MBKM) Penelitian/Riset di Badan Riset dan Inovasi Nasional (BRIN) KST Samaun Sadikun Bandung.

KATA INSPIRASI

"Maka sesungguhnya bersama kesulitan ada kemudahan, sesungguhnya bersama kesulitan ada kemudahan."
(QS. Al-Insyirah: 5–6)

"Katakanlah, sesungguhnya salatku, ibadahku, hidupku, dan matiku hanyalah untuk Allah, Tuhan semesta alam."
(QS. Al-An'am: 162)

"Ketika semua terasa mustahil, yakinlah ada Allah yang dapat membuat semuanya menjadi mungkin karena 'innamā amruhu idzā arāda syai'an ay yaqūla lahu kun fa yakūn'."
(QS. Yā-Sīn: 82)

"Lā ilāha illā anta, subḥānaka innī kuntu mina-ālimīn."
(Tidak ada Tuhan selain Engkau, Maha Suci Engkau, sesungguhnya aku termasuk orang-orang yang zalim)
(QS. Al-Anbiyā': 87)

"keep calm sebab segala sesuatu telah tertulis di Lauhul Mahfuz"

PERSEMBAHAN

Bismillāhirrahmānirrahīm, Alhamdulillāhirabbil ‘ālamīn, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas limpahan nikmat, rahmat, dan hidayah-Nya sehingga skripsi ini dapat diselesaikan dengan baik. Salawat serta salam semoga senantiasa tercurahkan kepada junjungan Nabi Muhammad SAW, suri teladan bagi seluruh umat manusia. Dengan penuh rasa syukur, penulis mempersembahkan karya ini sebagai ungkapan terima kasih kepada:

Ibunda Tercinta

Terima kasih kepada mama tercinta atas kasih sayang yang tak pernah putus, doa yang senantiasa mengiringi setiap langkah, serta dukungan dan penguatan yang selalu diberikan. Ketulusan, kesabaran, dan keikhlasan mama menjadi sumber kekuatan terbesar bagi penulis dalam menempuh perjalanan pendidikan ini.

Kedua Kakak Laki-laki Tercinta

Terima kasih kepada kedua cak-ku yang senantiasa berjuang dan bekerja keras demi pendidikan penulis, serta dengan penuh tanggung jawab menggantikan peran seorang ayah sejak penulis masih kecil. Pengorbanan, perhatian, dan usaha kalian dalam menyekolahkan penulis hingga akhirnya dapat menyelesaikan skripsi dan meraih gelar sarjana merupakan bentuk kasih dan pengorbanan yang sangat penulis syukuri dengan sepenuh hati.

Keluarga Tercinta

Terima kasih kepada seluruh anggota keluarga yang senantiasa memberikan dukungan, motivasi, doa, dan perhatian dalam berbagai bentuk yang mungkin tidak dapat penulis sebutkan satu per satu. Kehadiran, kepedulian, dan kehangatan keluarga menjadi kekuatan besar bagi penulis dalam menjalani setiap proses hingga terselesaikannya skripsi ini.

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan dosen pembahas atas bimbingan, arahan, kesabaran, serta ilmu yang sangat berharga selama proses penyusunan skripsi ini. Semoga Allah SWT senantiasa melimpahkan rahmat dan keberkahan-Nya atas segala kebaikan yang telah diberikan.

Sahabat-sahabatku

Terimakasih kepada semua orang-orang baik yang senantiasa hadir dalam suka maupun duka, saling membantu ketika menghadapi kesulitan, bertumbuh dan berjuang bersama, serta saling menguatkan dalam setiap proses yang dijalani. Semoga kebersamaan dan kebaikan ini menjadi berkah di dunia dan akhirat.

Almamater Tercinta
Universitas Lampung

SANWACANA

Alhamdulillah, puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul "*Analisis Pengaruh Latent Dirichlet Allocation dan Graph Convolutional Network pada Model X untuk Named Entity Recognition Berita Penyakit Menular Berbahasa Indonesia*" dengan baik dan lancar serta tepat pada waktu yang telah ditentukan. Shalawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad SAW.

Dalam proses penyusunan skripsi ini, banyak pihak yang telah membantu memberikan bimbingan, dukungan, arahan, motivasi serta saran sehingga skripsi ini dapat terselesaikan. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Ibu Dr. Dian Kurniasari, S.Si., M.Sc. selaku Pembimbing 1 yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, bantuan, motivasi, saran serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
2. Ibu Dr. Purnomo Husnul Khotimah, M.T. selaku Pembimbing II yang telah banyak meluangkan waktunya untuk memberikan arahan, bimbingan, bantuan, motivasi, saran serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
3. Ibu Prof. Dra. Wamiliana, MA., Ph.D. selaku Penguji yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat menjadi lebih baik lagi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Ibu Siti Laelatul Chasanah, S.Pd., M.Si. selaku dosen pembimbing akademik.

6. Seluruh dosen, staff dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Kedua orang tua tercinta, Terima kasih atas anugerah kehidupan serta kecerdasan yang telah diturunkan kepada penulis. Kepada mama, terima kasih atas seluruh kasih sayang, doa, dukungan, arahan, dan pengorbanan tanpa henti yang senantiasa membimbing penulis menuju jalan yang lebih baik.
8. Keluarga tercinta, Cak Lek, Cak Det, Yuk Ninit, Mba Rini, Yuk Yen, Kak Aan, serta keponakan-keponakan tersayang Kifa, Nuri, Dira, Zuya, dan Dapun. Terima kasih atas dukungan, kerja keras, pengorbanan, kesabaran, dan keikhlasan dalam membersamai penulis sebagai anak bungsu yang menyandang status yatim sejak kecil.
9. Prof. Dr. Nairobi, S.E., M.Si. dan Tante Andri. Terima kasih atas kasih sayang, perhatian, kebaikan, dan kepedulian selama penulis menempuh pendidikan di Universitas Lampung.
10. Sahabat-sahabat tercinta, Nazla, Ketut, Anita, Nisa, Tama, Fatur, dan Venny. Terima kasih atas kebersamaan, bantuan, serta dukungan dalam proses belajar dan penyusunan skripsi ini, terutama pada masa-masa sulit. Kehadiran dan penguatan kalian menjadi bagian penting hingga skripsi ini dapat diselesaikan dengan baik.
11. Jodoh penulis yang telah tertulis di Lauhul Mahfuz, yang menjadi salah satu motivasi penulis dalam menyelesaikan skripsi ini. Penulis meyakini bahwa segala sesuatu yang ditakdirkan akan menemukan jalannya pada waktu terbaik. Skripsi ini menjadi bukti kesungguhan penulis dalam mengutamakan pendidikan, bahwa dari awal perkuliahan hingga penyelesaian tugas akhir ini dijalani tanpa kehadiran laki-laki mana pun.

Semoga skripsi ini dapat bermanfaat bagi kita semua. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, sehingga penulis mengharapkan kritik dan saran yang membangun untuk menjadikan skripsi ini lebih baik lagi.

Bandar Lampung, 14 April 2026

Erin Elfitriani

DAFTAR ISI

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan teknologi digital telah mendorong perubahan dalam cara masyarakat mengakses dan menyebarkan informasi. Berbagai platform berita daring (*online news media*) kini menjadi salah satu rujukan utama masyarakat untuk memperoleh informasi terkini, termasuk yang berkaitan dengan kesehatan dan penyebaran penyakit menular. Ketika pandemi COVID-19 melanda, arus berita yang cepat dan masif terbukti memainkan peran penting dalam membentuk persepsi serta kesadaran masyarakat terhadap kondisi kesehatan global (Krawczyk dkk., 2021). Melalui pemberitaan tersebut, berbagai informasi mengenai perkembangan kasus, wilayah terdampak, dan kebijakan penanganan tersaji secara luas kepada publik. Penyajian informasi yang demikian menunjukkan bahwa data yang berasal dari berita daring memiliki potensi besar untuk dimanfaatkan dalam pemantauan dan mitigasi penyebaran penyakit. Meskipun demikian, pemanfaatan data tersebut menghadapi kendala karena sebagian besar masih berbentuk teks tidak terstruktur sehingga sulit diproses secara otomatis tanpa bantuan teknologi pemrosesan bahasa alami atau *Natural Language Processing* (NLP) (Rakhmawati dkk., 2024).

Salah satu teknik utama dalam NLP yang banyak digunakan untuk menata informasi dari teks tidak terstruktur adalah *Named Entity Recognition* (NER). Teknik ini berfungsi untuk mengenali dan mengelompokkan entitas penting dalam teks, seperti nama orang (*Person*), organisasi (*Organization*), dan lokasi (*Location*), termasuk entitas kontekstual lain yang relevan dengan bidang tertentu seperti istilah penyakit (*Disease*) yang sering muncul pada berita kesehatan (Jehangir dkk., 2023). Melalui kemampuan tersebut, NER membantu sistem komputer mengidentifikasi penyakit, wilayah terdampak, dan lembaga yang terlibat dalam penanganan. Informasi yang semula tersebar dalam bentuk teks dapat diubah menjadi data yang lebih terorganisasi sehingga mendukung analisis tren, pemetaan penyebaran penyakit,

hingga perumusan kebijakan berbasis data (De Magistris dkk., 2022).

Dalam implementasinya, penerapan NER pada teks berbahasa Indonesia masih menghadapi sejumlah tantangan. Struktur kalimat dalam berita daring tidak selalu seragam dan istilah yang digunakan sering kali bervariasi. Nama penyakit yang sama dapat muncul dengan bentuk berbeda seperti “Covid-19”, “corona virus”, atau “virus covid”. Perbedaan ini membuat model sulit mengenali entitas secara konsisten (Zainuddin & Tahir, 2025). Tantangan lain yang sering ditemukan adalah ketidakseimbangan distribusi label pada dataset, di mana sebagian besar token berlabel *Outside* (O), sedangkan entitas penting seperti *Disease* dan *Organization* hanya muncul dalam jumlah kecil. Kondisi tersebut membuat model cenderung fokus pada kelas dominan dan kurang mampu menangkap entitas minor yang justru berperan penting dalam konteks kesehatan masyarakat (Archana & Prakash, 2024).

Karakteristik judul berita daring turut memperkuat tantangan tersebut. Berbeda dari isi berita penuh yang memiliki narasi panjang, judul bersifat sangat ringkas, langsung menyampaikan inti peristiwa, dan sering kali memuat beberapa entitas penting dalam ruang teks yang terbatas (Krawczyk dkk., 2021). Keterbatasan konteks ini menjadikan tugas NER semakin menantang karena model harus mampu mengenali entitas dengan tepat tanpa banyak informasi pendukung. Kondisi ini mendorong perlunya pendekatan pemodelan yang lebih mendalam dan adaptif agar sistem dapat bekerja efektif meskipun informasi yang tersedia singkat dan padat.

Sejumlah penelitian sebelumnya menunjukkan bahwa model NER yang diperkuat dengan kombinasi berbagai teknik representasi mampu meningkatkan kinerja model. Penelitian yang dilakukan Zainuddin dan Tahir (2025) menunjukkan bahwa pendekatan *hybrid* berbasis Transformer, Word2Vec, dan Bi-LSTM mampu menangkap konteks sekaligus urutan kata, menghasilkan peningkatan performa yang konsisten pada teks berita berbahasa Indonesia. Temuan oleh Umam dkk. (2025a) menegaskan pentingnya pemerkayaan konteks dengan memanfaatkan *Latent Dirichlet Allocation* (LDA) untuk mendukung proses praanotasi, yang terbukti membantu model menangani keragaman istilah pada laporan pengaduan masyarakat berskala besar. Di sisi lain, Chen dan Shen (2025) menemukan bahwa mengintegrasikan struktur dependensi dan *Graph Convolutional Network* (GCN) dapat meningkatkan representasi fitur, sehingga relasi antartoken menjadi lebih jelas meskipun berada pada kalimat yang singkat. Tantangan ketidakseimbangan label juga menjadi perhatian Archana dan Prakash (2024), yang menunjukkan

bahwa pendekatan *undersampling* seperti iBUS mampu membantu model lebih peka terhadap entitas minor. Selain itu, De Magistris dkk. (2022) menunjukkan bahwa pemanfaatan NER dalam konteks analisis berita kesehatan dan deteksi *fake news* dapat menghasilkan informasi yang lebih akurat ketika entitas dikenali secara tepat.

Meskipun berbagai pendekatan tersebut telah menunjukkan peningkatan kinerja pada model NER, sebagian besar penelitian masih berfokus pada penerapan metode secara terpisah atau pada jenis teks yang memiliki konteks relatif lebih panjang. Penelitian yang secara khusus mengkaji penerapan model NER *hybrid* pada judul berita daring berbahasa Indonesia, dengan mempertimbangkan keterbatasan konteks, ketidakseimbangan label, serta integrasi pemodelan relasi antartoken, masih relatif terbatas. Oleh karena itu, penelitian ini menggunakan model X, yaitu model *hybrid* yang menggabungkan IndoBERT dan BiLSTM untuk menangkap informasi berbasis konteks dan urutan kata pada judul berita daring. Selanjutnya, penelitian ini menganalisis pengaruh penambahan LDA dan GCN sebagai upaya untuk mengatasi tantangan yang telah diuraikan sebelumnya. Analisis dilakukan melalui empat skenario, yaitu model X sebagai *baseline*, model X dengan LDA, model X dengan GCN, serta model X dengan LDA dan GCN. Untuk memperoleh kinerja optimal pada masing-masing skenario, penentuan konfigurasi terbaik dilakukan melalui *hyperparameter tuning* menggunakan Optuna, sehingga model memperoleh pengaturan parameter yang paling sesuai sebelum pelatihan akhir dilakukan.

Penelitian ini diharapkan memberikan kontribusi bagi pengembangan sistem NER berbahasa Indonesia, khususnya di bidang kesehatan. Hasil penelitian dapat dimanfaatkan untuk analisis berita daring, pemantauan penyebaran penyakit, serta mendukung penelitian lanjutan dalam bidang ekstraksi informasi dan analisis teks. Pendekatan ini diharapkan turut mendorong pemanfaatan teknologi NLP untuk mendukung sistem deteksi dini dan pengelolaan informasi kesehatan di Indonesia secara lebih cepat, akurat, dan efisien.

Penelitian ini disusun menjadi lima bagian. Bagian 1 menyajikan latar belakang, rumusan masalah, tujuan penelitian, dan manfaat penelitian. Bagian 2 berisi kajian terdahulu dan literatur yang berkaitan dengan penelitian. Bagian 3 menjelaskan metode dan kerangka kerja penelitian. Bagian 4 membahas serta menganalisis hasil penelitian secara mendalam. Terakhir, Bagian 5 menyajikan kesimpulan penelitian.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, penelitian ini dilaksanakan untuk menjawab beberapa permasalahan yang muncul dalam proses pengenalan entitas bernama (*Named Entity Recognition*) pada teks berita berbahasa Indonesia, khususnya pada domain kesehatan. Permasalahan tersebut dapat dirumuskan sebagai berikut:

1. Bagaimana penerapan model X, yaitu model *Hybrid* IndoBERT dan BiLSTM dalam melakukan NER pada judul berita daring mengenai penyakit menular berbahasa Indonesia?
2. Bagaimana pengaruh penambahan LDA dan GCN pada model X dalam mengatasi tantangan pengenalan entitas, khususnya yang berkaitan dengan ketidakseimbangan label, keterbatasan konteks, dan relasi antartoken?
3. Sejauh mana peningkatan kinerja model X dengan penambahan LDA dan GCN dibandingkan dengan model X sebagai *baseline* berdasarkan hasil pengukuran *accuracy*, *precision*, *recall*, dan *F1-score*?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengembangkan model X, yaitu model *hybrid* IndoBERT dan BiLSTM, untuk melakukan NER pada judul daring berbahasa Indonesia terkait penyakit menular.
2. Menganalisis pengaruh penambahan LDA dan GCN pada model X dalam mengatasi tantangan pengenalan entitas, khususnya ketidakseimbangan label, keterbatasan konteks, dan relasi antartoken.
3. Mengevaluasi peningkatan kinerja model X dengan penambahan LDA dan GCN menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengetahui efektivitas pendekatan dibandingkan model X sebagai *baseline*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan beberapa manfaat sebagai berikut:

1. Memberikan kontribusi ilmiah dalam bidang *Natural Language Processing* (NLP), khususnya pada pengembangan sistem *Named Entity Recognition* berbahasa Indonesia di domain kesehatan.
2. Menyediakan referensi mengenai penerapan model X sebagai model *hybrid* berbasis IndoBERT dan BiLSTM, serta pengaruh penambahan LDA dan GCN terhadap kinerja model NER, sehingga dapat diadaptasi untuk analisis data berita daring berbahasa Indonesia di bidang lainnya.
3. Menjadi acuan dalam pengembangan sistem analisis berita daring dan pemantauan isu kesehatan secara otomatis untuk mendukung deteksi dini dan pemetaan penyebaran penyakit menular di Indonesia.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian ini merujuk pada sejumlah studi terdahulu yang digunakan sebagai acuan dalam mengidentifikasi kelemahan model sebelumnya serta menentukan celah penelitian yang dapat diisi. Penelitian-penelitian tersebut memberikan dasar teoretis dan empiris untuk mendukung pengembangan model yang diusulkan. Daftar penelitian terkait disajikan pada Tabel 1.

Tabel 1. Penelitian Terkait

No.	**Penelitian**	**Data**	**Metode**	**Hasil (%)**		
				Prec.	**Rec.**	**F1**
1	(Zainuddin & Tahir, 2025) – *Entity Extraction in Indonesian Online News Using Named Entity Recognition (NER) with Hybrid Method Transformer, Word2Vec, Attention and Bi-LSTM*	Dataset berita daring berbahasa Indonesia	Hybrid Transformer + Word2Vec + Bi-LSTM	86.71	83.64	**85.11**
2	(Umam et al., 2025a) – *Enhancing Entity Extraction in E-Government Complaint Data Using LDA-Assisted NER*	53,858 laporan LaporGub (2022–2025)	spaCy NER + LDA (*pre-annotation*)	90.03	81.86	**85.75**
3	(Chen & Shen, 2025) – *NER Based on Dependency Structure Feature Fusion*	Dataset Catalan (SemEval 2010 Task 1)	Dependency + GCN + BiLSTM + CRF	84.36	79.48	**81.85**

No.	Penelitian	Data	Metode	Hasil (%)		
				Prec	Rec	F1
4	(Archana & Prakash, 2024) – *Biomedical Named Entity Recognition Through Improved Balanced Undersampling for Addressing Class Imbalance and Preserving Contextual Information*	Dataset NCBI Disease	CRF + iBUS (*undersampling*)	76.64	75.52	**76.08**
5	(De Magistris et al., 2022) – *An Explainable Fake News Detector Based on NER and Stance Classification Applied to COVID-19*	Dataset FNC-1 (train) & COVID19 -FNIR (eval)	NER + Stance Classification (XAI)	–	–	–

Penjelasan dari penelitian pada Tabel 1 sebagai berikut:

1. Penelitian Pertama (Zainuddin & Tahir, 2025)

   Zainuddin dan Tahir (2025) meneliti ekstraksi entitas pada berita daring berbahasa Indonesia menggunakan model *Hybrid* Transformer, Word2Vec, Attention, dan BiLSTM (TWBiL). Penelitian ini dilatarbelakangi oleh tantangan ambiguitas makna dan fleksibilitas struktur kalimat bahasa Indonesia yang sering menyebabkan kesalahan dalam pengenalan entitas. Model TWBiL menggabungkan *contextual embedding* berbasis Transformer, *semantic similarity* dari Word2Vec, serta kemampuan *sequence learning* dari BiLSTM dengan dukungan mekanisme *Attention*.

   Dataset yang digunakan berupa teks berita daring dengan entitas *Person*, *Location*, dan *Organization*. Hasil pengujian menunjukkan bahwa TWBiL mencapai *precision* 86,71%, *recall* 83,64%, dan *F1-score* 85,11%, ebih tinggi dibandingkan model BiLSTM tunggal (F1 75,18%). Penelitian ini menegaskan bahwa kombinasi *embedding* hibrida mampu menangkap konteks semantik yang lebih kaya dan menurunkan kesalahan identifikasi entitas ambigu, meskipun memerlukan waktu pelatihan yang lebih besar.

2. Penelitian Kedua (Umam dkk., 2025a)

   Umam dkk. (2025a) mengusulkan pendekatan *LDA-Assisted NER* untuk meningkatkan akurasi ekstraksi entitas pada data teks laporan publik Indonesia.

Data penelitian diambil dari platform LaporGub, yang berisi 53.858 laporan masyarakat periode 2022–2025. Sebelum dilakukan pelabelan, penulis menerapkan *Latent Dirichlet Allocation* (LDA) untuk mendeteksi topik dominan pada teks dan memperkaya konteks semantik dalam proses anotasi entitas.

Model NER dikembangkan menggunakan pustaka *spaCy* dengan tiga label utama, yaitu *Person*, *Organization*, dan *Location*. Hasil evaluasi menunjukkan bahwa pendekatan ini memberikan peningkatan performa dibandingkan model *baseline*, dengan *precision* 90,03%, *recall* 81,86%, dan *F1-score* 85,75%. Integrasi LDA terbukti membantu meningkatkan konsistensi anotasi, terutama untuk entitas yang kontekstual seperti nama lembaga dan lokasi. Penelitian ini menekankan bahwa pendekatan tematik semantik seperti LDA dapat meningkatkan kualitas model NER tanpa kebutuhan komputasi besar seperti Transformer.

3. Penelitian Ketiga (Chen & Shen, 2025)

Chen dan Shen (2025) mengembangkan model NER berbasis *dependency structure feature fusion* untuk meningkatkan kemampuan model dalam memahami hubungan sintaktik antartoken. Peneliti menggabungkan *Graph Convolutional Network* (GCN) dengan *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Conditional Random Field* (CRF) agar model dapat memanfaatkan baik konteks urutan kata maupun struktur dependensi dalam kalimat.

Eksperimen dilakukan pada dataset SemEval 2010 Task 1 yang menggunakan bahasa Catalan, sebuah bahasa yang digunakan di wilayah Catalonia (Spanyol), Andorra, serta beberapa bagian dari Prancis dan Kepulauan Balearic. Hasil pengujian menunjukkan bahwa model GCN-BiLSTM-CRF mencapai *precision* 84,36%, *recall* 79,48%, dan *F1-score* 81,85%, lebih tinggi dibandingkan BiLSTM-CRF standar yang hanya memperoleh F1 77,10%. Penelitian ini menegaskan bahwa representasi berbasis graf melalui GCN efektif dalam menangkap hubungan sintaktik dan dependensi panjang antar-token, sehingga meningkatkan kinerja pengenalan entitas multibahasa.

4. Penelitian Keempat (Archana & Prakash, 2024)

   Archana dan Prakash (2024) meneliti masalah ketidakseimbangan kelas dalam pengenalan entitas biomedis dan mengusulkan pendekatan *Improved Balanced Undersampling* (iBUS) untuk mempertahankan keseimbangan distribusi label sambil menjaga konteks semantik. Model dikembangkan berbasis CRF dan diuji pada dataset NCBI Disease serta dataset biomedis lainnya.

   Hasil eksperimen menunjukkan bahwa pendekatan iBUS berhasil meningkatkan kinerja model, dengan *precision* 76,64%, *recall* 75,52%, dan *F1-score* 76,08%. Pendekatan ini secara konsisten menghasilkan peningkatan performa dibanding metode pembelajaran standar tanpa *balancing*, serta menunjukkan keunggulan dalam mendeteksi entitas minor tanpa mengorbankan akurasi keseluruhan. Penelitian ini menegaskan pentingnya strategi penyeimbangan data yang mempertahankan konteks semantik agar model tidak bias terhadap kelas dominan, khususnya di domain biomedis yang memiliki entitas langka.

5. Penelitian Kelima (De Magistris dkk., 2022)

   De Magistris dkk. (2022) mengembangkan sistem deteksi berita palsu berbasis *Explainable Artificial Intelligence* (XAI) yang menggabungkan NER dan *Stance Classification* untuk mendeteksi berita palsu terkait pandemi COVID-19. Dataset yang digunakan terdiri dari lebih dari 1,6 juta artikel berita global yang dikumpulkan antara tahun 2019 hingga 2020 dari berbagai sumber, termasuk BBC News dan AYLIEN COVID-19 News Dataset.

   Sistem ini menggunakan modul NER untuk mengekstraksi entitas penting seperti nama tokoh, organisasi, dan lokasi yang sering muncul dalam berita palsu, kemudian melakukan *stance classification* untuk menentukan posisi teks terhadap klaim tertentu. Hasil penelitian menunjukkan bahwa integrasi kedua teknik ini meningkatkan interpretabilitas sistem deteksi berita palsu dan mampu mengidentifikasi sumber misinformasi dengan lebih akurat. Pendekatan ini menegaskan pentingnya NER dalam mendukung sistem deteksi berita berbasis entitas, terutama dalam konteks pandemi yang penuh dengan informasi berlebihan (*infodemic*).

Seluruh penelitian terdahulu secara umum berfokus pada upaya peningkatan kinerja NER melalui dua pendekatan utama, yaitu pengayaan konteks dan pemodelan relasi antartoken. Pendekatan pengayaan konteks dilakukan dengan memanfaatkan model representasi semantik berbasis embedding, seperti Transformer dan Word2Vec, untuk menangkap makna kontekstual kata dalam kalimat. Sementara itu, LDA digunakan secara terpisah sebagai metode *topic modeling* guna mengekstraksi konteks global pada tingkat dokumen sebagai informasi pendukung. Adapun pemodelan relasi diwujudkan melalui arsitektur berbasis urutan dan graf, seperti BiLSTM, GCN, dan CRF. Kedua pendekatan tersebut bertujuan untuk meningkatkan kemampuan model dalam mengenali entitas pada teks tidak terstruktur dengan mempertimbangkan makna semantik dan keterkaitan sintaksis antarkata. Perkembangan ini menunjukkan pergeseran dari model statistik konvensional menuju model *deep learning* yang kontekstual dan adaptif terhadap kompleksitas bahasa alami.

2.2 Berita Daring

Berita daring (*online news*) merupakan media informasi digital yang berkembang pesat seiring meningkatnya penggunaan internet dan perangkat mobile. Media berita digital memiliki karakteristik penyebaran informasi yang cepat, jangkauan luas, serta kemampuan memperbarui konten secara *real time*. Dalam konteks kesehatan publik, keberadaan berita daring menjadi sangat penting karena masyarakat menjadikannya sebagai sumber utama untuk memperoleh informasi terkini mengenai wabah, kebijakan kesehatan, dan risiko penyakit (Mach dkk., 2021).

Pada masa pandemi COVID-19, berita daring berperan besar dalam membentuk persepsi risiko masyarakat. Mach dkk. (2021) menemukan bahwa liputan media digital memiliki dua karakteristik dominan, yaitu tingkat kualitas ilmiah dan unsur sensasionalisme. Mayoritas media menyajikan berita dengan kualitas ilmiah menengah dan sensasionalisme rendah, sehingga mampu membantu mengomunikasikan risiko secara lebih seimbang. Namun demikian, bias politik dan kebijakan redaksi tetap memengaruhi cara media menyusun judul serta membingkai informasi kesehatan. Berita daring juga menjadi ruang yang rentan terhadap penyebaran misinformasi. Rocha dkk. (2023) menunjukkan bahwa selama masa pandemi, banjir informasi yang tidak tervalidasi memicu peningkatan kecemasan dan kepanikan publik. Informasi yang salah mengenai obat-obatan, pencegahan penyakit, dan teori konspirasi menyebar cepat melalui kanal digital sehingga berdampak pada

perilaku kesehatan masyarakat secara luas.

Aspek lain yang semakin menonjol dalam berita daring adalah meningkatnya ujaran kebencian terhadap kelompok tertentu, terutama selama situasi krisis kesehatan. Castaño-Pulgarín dkk. (2021) menemukan bahwa pandemi COVID-19 menyebabkan peningkatan signifikan ujaran kebencian di ruang digital yang dipicu oleh kecemasan dan ketidakpastian publik terhadap asal-usul serta penyebaran penyakit. Fenomena ini menggambarkan bahwa media daring bukan hanya menjadi saluran informasi kesehatan, tetapi juga arena munculnya reaksi sosial negatif. Dengan demikian, berita daring memiliki peran ganda dalam kesehatan masyarakat yaitu sebagai sumber informasi yang diperlukan publik, tetapi sekaligus berpotensi memperbesar penyebaran misinformasi dan dampak psikososial. Kondisi ini menegaskan pentingnya analisis terhadap judul berita daring, karena struktur judul, pemilihan diksi, dan gaya penyajian dapat memengaruhi persepsi masyarakat terhadap risiko kesehatan.

2.3 Penyakit Menular

Penyakit menular merupakan gangguan kesehatan yang disebabkan oleh mikroorganisme seperti virus, bakteri, dan parasit yang dapat berpindah dari satu individu ke individu lainnya. Penularan dapat terjadi melalui udara, droplet, kontak langsung, maupun melalui media perantara. Dinamika penyebaran sangat dipengaruhi oleh faktor lingkungan, mobilitas penduduk, serta pola interaksi sosial yang memungkinkan agen penyakit menyebar lebih cepat antarindividu (Sarantopoulos dkk., 2024).

Dalam kajian epidemiologi modern, penyakit menular tidak hanya dipahami dari sisi biologis, tetapi juga dari aspek sosial. Ketidakpastian informasi, persepsi risiko, dan respons masyarakat memiliki peran besar dalam membentuk pola penyebaran penyakit. Menurut Bin Naeem dan Boulos (2021), pemahaman masyarakat terhadap penyakit sangat dipengaruhi oleh informasi yang mereka terima, terutama ketika berada dalam situasi wabah. Ketika informasi tidak lengkap atau tidak akurat, masyarakat dapat mengalami kebingungan yang berdampak pada munculnya perilaku kesehatan yang kurang tepat. Oleh karena itu, penyediaan informasiyang jelas dan terpercaya menjadi elemen penting dalam upaya pengendalian penyakit menular.

Pandemi COVID-19 menjadi salah satu contoh paling jelas mengenai bagaimana penyakit menular modern dapat memberikan dampak luas bagi aspek klinis maupun sosial. COVID-19 disebabkan oleh virus SARS-CoV-2 yang memiliki kemampuan penyebaran yang sangat cepat, termasuk melalui individu yang belum menunjukkan gejala. Excler dkk. (2021) menjelaskan bahwa penularan pra-gejala menjadi salah satu alasan utama mengapa COVID-19 menyebar dengan cepat pada fase awal kemunculannya. Kondisi ini membuat proses identifikasi dan isolasi kasus menjadi lebih menantang dibandingkan penyakit menular konvensional.

Secara klinis, COVID-19 memiliki ragam gejala yang luas, mulai dari keluhan ringan seperti demam dan batuk hingga kondisi berat seperti pneumonia dan gangguan pernapasan akut. Baker dkk. (2022) menunjukkan bahwa tingkat keparahan penyakit berkaitan erat dengan usia pasien, adanya komorbiditas, serta respons tubuh terhadap infeksi. Gejala yang tidak seragam membuat proses diagnosis menjadi lebih kompleks, terutama di wilayah dengan keterbatasan fasilitas kesehatan.

Dari perspektif penularan, COVID-19 memiliki pola penyebaran yang dipengaruhi oleh beban virus dan intensitas kontak antarindividu. Excler dkk. (2021), menegaskan bahwa variasi beban virus pada setiap individu dapat menyebabkan terjadinya *super-spreading events*, yaitu kondisi ketika satu orang dapat menularkan virus ke banyak orang dalam waktu singkat. Pola penyebaran semacam ini mempercepat perluasan wabah di berbagai wilayah, terutama di daerah padat penduduk.

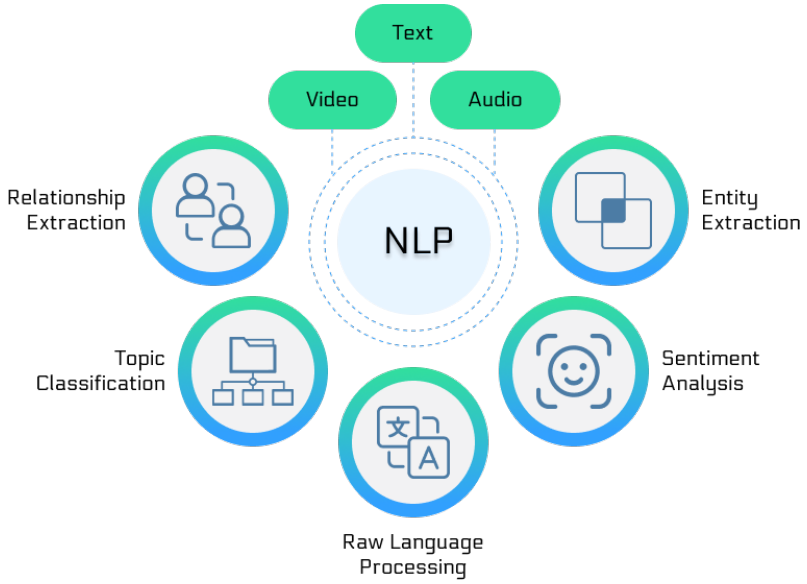
Selain aspek medis, pandemi COVID-19 turut memengaruhi keadaan psikologis masyarakat. Ju dkk. (2023) menemukan bahwa masyarakat mengalami peningkatan kecemasan dan tekanan emosional akibat paparan informasi mengenai penyakit yang berlangsung terus-menerus. Hal ini diperparah oleh beredarnya misinformasi di berbagai media daring yang membuat masyarakat semakin sulit membedakan informasi yang benar atau keliru. Akibatnya, pandemi tidak hanya menjadi peristiwa kesehatan, tetapi juga peristiwa sosial yang memengaruhi perilaku, persepsi, dan kondisi psikologis masyarakat.

Dengan demikian, Pembahasan mengenai penyakit menular tidak dapat dipisahkan dari dinamika penyebaran COVID-19 karena pandemi tersebut memperlihatkan bagaimana interaksi antara faktor biologis, sosial, dan lingkungan membentuk pola transmisi penyakit dalam skala luas. COVID-19 menunjukkan bahwa penyebaran

penyakit modern dipengaruhi bukan hanya oleh karakteristik patogen, tetapi juga oleh respons masyarakat terhadap informasi kesehatan yang beredar di ruang publik. Sejumlah penelitian menegaskan bahwa persepsi risiko, kecemasan, dan kepatuhan terhadap protokol kesehatan berkaitan erat dengan kualitas informasi yang diterima masyarakat melalui berbagai media digital. Literatur mengenai penyakit menular dan COVID-19 memberikan landasan teoretis yang penting dalam memahami keterkaitan antara penyebaran penyakit dan arus informasi kesehatan di media daring.

2.4 *Natural Language Processing* (NLP)

NLP merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada interaksi antara komputer dan bahasa manusia. Tujuan utama NLP adalah membuat mesin mampu memahami, menafsirkan, dan menghasilkan bahasa alami sehingga dapat dimanfaatkan dalam berbagai aplikasi seperti penerjemahan otomatis, analisis sentimen, sistem tanya jawab, dan ekstraksi informasi. Dalam penelitian berbasis teks, NLP berperan mengubah data tidak terstruktur menjadi informasi terformat yang dapat diproses secara komputasional untuk menghasilkan pemahaman yang lebih sistematis (Amien, 2023).



Gambar 1. *Natural Language Processing* (Amazinum, 2023)

Secara umum, proses NLP mencakup beberapa tahapan utama yaitu *text preprocessing*, representasi kata melalui *feature extraction* atau *embedding*,

pemodelan konteks menggunakan algoritma *machine learning* atau *deep learning*, hingga tahap keluaran berupa analisis seperti klasifikasi topik, analisis sentimen, atau ekstraksi entitas. Menurut Amazinum (2023), sistem NLP bekerja dengan menerima masukan berupa teks, suara, atau video yang kemudian diproses menjadi bentuk yang dapat dipahami oleh mesin. Hasil pemrosesan tersebut memungkinkan sistem melakukan berbagai fungsi seperti *topic classification*, *relationship extraction*, *entity extraction*, dan *sentiment analysis* yang direpresentasikan pada Gambar 1.

NLP telah digunakan secara luas di berbagai sektor, mulai dari layanan publik, hukum, kesehatan, hingga media daring. Subowo dkk. (2025) menyebutkan bahwa dalam bidang hukum, NLP dimanfaatkan untuk mengenali entitas hukum dari teks putusan pengadilan. Dalam konteks pemerintahan digital, Umam dkk. (2025a) mengimplementasikan NLP untuk mengekstraksi entitas penting dari laporan masyarakat di platform *e-government*, seperti nama lembaga, lokasi, dan jenis pengaduan. Sementara itu, Nuryanto (2025) menunjukkan bahwa penerapan model IndoBERT berbasis NLP dapat digunakan untuk mendeteksi berita hoaks berbahasa Indonesia karena mampu memahami konteks bahasa secara lebih akurat. Temuan ini membuktikan bahwa NLP memiliki potensi besar dalam mendukung analisis informasi lintas bidang, termasuk kesehatan digital dan isu publik.

Kemajuan teknologi *deep learning* dan model Transformer seperti BERT serta IndoBERT telah meningkatkan kemampuan NLP dalam memahami konteks kalimat yang kompleks. Bahasa Indonesia sendiri memiliki struktur morfologi yang kaya, konteks lokal yang beragam, dan pola susunan kalimat yang relatif fleksibel. Amien dan Gunawan (2024) menjelaskan bahwa karakteristik tersebut membuat model NLP umum yang dilatih menggunakan bahasa global tidak selalu optimal ketika diterapkan secara langsung pada Bahasa Indonesia. Selain itu, keterbatasan dataset beranotasi dan variasi gaya penulisan antara teks formal dan informal semakin menambah tantangan dalam pengembangan NLP Indonesia (Amien, 2023).

Perkembangan teknologi NLP dapat dibagi menjadi beberapa generasi pendekatan. Pendekatan awal menggunakan sistem berbasis aturan (*rule-based system*) dengan pola linguistik manual, namun metode ini terbatas dalam menangani variasi bahasa. Selanjutnya muncul pendekatan *machine learning* seperti *Hidden Markov Model* (HMM) dan CRF yang mampu mengenali pola dari data beranotasi. Evolusi berikutnya adalah pendekatan *deep learning* menggunakan arsitektur *Recurrent Neural Network* (RNN), BiLSTM, dan Transformer seperti BERT (*Bidirectional*

Encoder Representations from Transformers). Model berbasis Transformer terbukti unggul karena mampu memahami konteks dua arah dan menangkap hubungan semantik antarkata dengan lebih baik (Amien & Gunawan, 2024).

Salah satu bentuk adaptasi penting dari model berbasis Transformer untuk Bahasa Indonesia adalah IndoBERT. Model ini dikembangkan dengan melatih BERT menggunakan korpus Bahasa Indonesia seperti Wikipedia, OSCAR, dan Kompas agar dapat menangkap karakteristik morfologis dan konteks lokal secara lebih akurat (Amien & Gunawan, 2024). Subowo dkk. (2025) juga menekankan bahwa penggunaan model yang dilatih pada data Indonesia sangat penting agar sistem mampu mengenali entitas sesuai konteks budaya dan linguistik Indonesia.

NLP memiliki peran penting sebagai fondasi utama untuk membangun sistem NER. NLP memungkinkan proses pengenalan entitas dilakukan secara otomatis melalui analisis konteks kalimat dan makna semantik. Penelitian oleh Rakhmawati dkk. (2024) menunjukkan bahwa penerapan LDA dalam tahap prapemrosesan mampu memperkaya representasi semantik sebelum pelatihan NER, sehingga meningkatkan efisiensi klasifikasi entitas. Secara keseluruhan, berbagai penelitian menunjukkan bahwa NLP berperan sebagai fondasi utama dalam memahami struktur, konteks, dan makna dalam data teks. Perkembangan metode dan model yang semakin canggih memungkinkan analisis yang lebih mendalam terhadap beragam bentuk informasi tekstual. Literatur ini memberikan pemahaman komprehensif mengenai kemampuan dan tantangan NLP pada berbagai domain, termasuk bidang kesehatan dan informasi publik.

2.5 *Named Entity Recognition* (NER)

NER merupakan salah satu tugas utama dalam NLP yang berfungsi untuk mengidentifikasi dan mengelompokkan entitas penting dari teks tidak terstruktur menjadi data yang terorganisir. Entitas yang umum dikenali meliputi nama orang (*Person*), organisasi (*Organization*), lokasi (*Location*), serta entitas lain sesuai domain tertentu seperti penyakit (*Disease*) dan obat (*Drug*) (Jehangir dkk., 2023). NER memiliki peran penting dalam sistem *information extraction* karena memungkinkan mesin memahami makna teks dengan cara yang menyerupai penalaran manusia, sehingga dapat mendukung berbagai aplikasi seperti analisis berita, biomedis, hukum, dan sistem rekomendasi informasi.

Metode NER mengalami perkembangan pesat dari pendekatan berbasis aturan (*rule-based system*) menuju model pembelajaran mendalam (*deep learning*). Pendekatan awal menggunakan pola linguistik manual memiliki keterbatasan dalam menangani variasi bahasa dan konteks kalimat. Pendekatan berikutnya menggunakan model statistik seperti HMM dan CRF yang mampu mengenali pola pada data beranotasi secara lebih konsisten (Jehangir dkk., 2023). Selanjutnya, muncul arsitektur *deep learning* seperti BiLSTM dan Transformer yang dapat memahami konteks kalimat secara dua arah serta menangkap relasi semantik antarkata dengan lebih mendalam (Amien & Gunawan, 2024).

Dalam bidang biomedis, NER digunakan untuk mengenali istilah seperti gen, penyakit, dan obat, yang kemudian dinormalisasi ke dalam terminologi baku seperti ICD-10 atau RxNorm. Tantangan utama NER biomedis terletak pada banyaknya istilah kompleks yang menyebabkan pelabelan manual sulit dilakukan. Noh dan Kavuluru (2021) mengusulkan pendekatan *joint learning* antara NER dan *entity normalization* untuk mengurangi kesalahan propagasi antartugas, dan hasilnya menunjukkan peningkatan performa deteksi entitas medis secara signifikan.

Selain di bidang medis, NER juga diterapkan di sektor hukum. Yulianti dkk. (2024) mengembangkan tugas *Legal Entity Recognition* (LER) menggunakan dataset IndoLER yang berisi dokumen keputusan pengadilan dengan 20 jenis entitas hukum seperti hakim, jaksa, dan advokat. Eksperimen mereka menunjukkan bahwa model berbasis Transformer seperti IndoRoBERTa dan XLM-RoBERTa lebih unggul dibandingkan BiLSTM-CRF dengan peningkatan F1-score hingga 7,9%. Hasil ini membuktikan bahwa model Transformer efektif dalam memahami konteks panjang dan struktur kalimat kompleks pada teks hukum berbahasa Indonesia.

Penerapan NER di bidang berita daring juga berkembang pesat. Zainuddin dan Tahir (2025) menggabungkan Transformer, Word2Vec, Attention, dan BiLSTM dalam model TWBiL untuk ekstraksi entitas pada berita daring berbahasa Indonesia. Model ini berhasil meningkatkan F1-score menjadi 85,11% dibandingkan BiLSTM tunggal yang hanya mencapai 75,18%. Penelitian ini menunjukkan bahwa penggabungan representasi semantik dan sekuensial dapat meningkatkan akurasi pengenalan entitas pada teks berita.

Khusus untuk Bahasa Indonesia, tantangan terbesar dalam tugas NER adalah keterbatasan dataset publik, kompleksitas morfologi, serta variasi gaya penulisan antar domain (Khairunnisa dkk., 2023). Bahasa Indonesia tergolong *low-resource language*, sehingga model yang efektif untuk bahasa global sering kali kurang optimal. Penelitian tersebut membuktikan bahwa penggunaan model BiLSTM–CRF dan Transformer seperti IndoBERT serta XLM-RoBERTa dapat meningkatkan konsistensi dan akurasi ekstraksi entitas setelah dilakukan *re-annotation* terhadap dataset publik.

Dalam domain biomedis, Archana dan Prakash (2024) menyoroti permasalahan ketidakseimbangan label entitas dan mengusulkan metode iBUS untuk menyeimbangkan distribusi data tanpa kehilangan konteks linguistik. Pendekatan ini meningkatkan F1-score sebesar 3,6% dibandingkan BiLSTM–CRF. Selain itu, Nemoto dkk. (2024) menunjukkan bahwa penyesuaian bobot label selama pelatihan dapat meningkatkan sensitivitas model terhadap entitas minor seperti *Disease* dan *Organization*.

Pendekatan semantik turut digunakan untuk memperkaya konteks antarentitas. Umam dkk. (2025a) mengintegrasikan LDA pada tahap prapemrosesan sebelum pelatihan model NER dan berhasil meningkatkan akurasi klasifikasi entitas pada data laporan publik. Hasil tersebut menunjukkan bahwa pengayaan konteks semantik efektif dalam membantu model memahami keterkaitan antarentitas yang jarang muncul.

Secara keseluruhan, penelitian mengenai NER menunjukkan tren menuju model berbasis *deep learning* dan Transformer yang dikombinasikan dengan strategi semantik serta penyeimbangan label. Tantangan utama yang masih dihadapi adalah bagaimana mengatasi ketidakseimbangan data antarentitas tanpa mengorbankan konteks semantik dan akurasi model.

2.6 Ketidakseimbangan label dalam Dataset NER

Ketidakseimbangan distribusi label (*class imbalance*) merupakan salah satu tantangan utama dalam pengembangan model NER. Kondisi ini muncul ketika jumlah sampel antarkelas berbeda cukup jauh. Pada tugas NER, sebagian besar token dalam korpus biasanya diberi label *Outside* (O) karena tidak termasuk dalam

kategori entitas apa pun, sementara entitas penting seperti *Disease*, *Organization*, atau *Location* hanya muncul dalam proporsi yang kecil. Ketimpangan ini membuat model cenderung belajar lebih banyak dari kelas dominan dan mengabaikan kelas minor, sehingga kemampuan sistem dalam mengenali entitas penting menjadi menurun (Archana & Prakash, 2024).

Ketidakseimbangan label dapat menurunkan kinerja model dan menimbulkan bias prediksi terhadap kelas mayoritas. Nemoto dkk. (2024) menemukan bahwa model NER yang dilatih pada data tidak seimbang lebih sering menghasilkan prediksi pada label dominan, terutama ketika proporsi token non-entitas jauh lebih besar dibandingkan token entitas. Untuk mengatasi hal tersebut, mereka mengusulkan pendekatan *Majority-or-Minority Learning* dengan memberi bobot pelatihan yang lebih besar pada kelas minor agar model tetap peka terhadap entitas langka tanpa menurunkan akurasi keseluruhan.

Permasalahan ketidakseimbangan label juga berhubungan dengan konteks linguistik dan struktur kalimat. Archana dan Prakash (2024) menekankan bahwa proses penyeimbangan data tidak boleh menghilangkan kalimat yang memuat informasi penting. Mereka mengembangkan metode iBUS yang menyeleksi data secara proporsional sambil mempertahankan konteks kalimat sehingga representasi semantik tetap terjaga. Hasil penelitian mereka menunjukkan peningkatan nilai F1-score pada tugas Biomedical NER karena metode ini membuat model tetap memahami hubungan antartoken meskipun jumlah data kelas minor diperbanyak.

Pada skala dokumen yang lebih besar, Lopez dkk. (2021) menemukan bahwa ketimpangan label membuat model sulit menangkap hubungan antarparagraf yang memuat entitas relevan. Dalam penelitian *document-level NER*, mereka menjelaskan bahwa rasio ekstrem seperti satu entitas positif di antara ribuan token negatif menyebabkan sistem gagal memahami konteks lintas kalimat. Penggunaan representasi berbasis konteks dokumen membantu meningkatkan akurasi dan *recall* untuk entitas minor yang jarang muncul.

Selain masalah proporsi data, ketidakseimbangan label juga memperkuat bias kontekstual dalam model. Noh dan Kavuluru (2021) menjelaskan bahwa ketika entitas muncul berulang dengan pola kalimat yang serupa, model cenderung mempelajari urutan kata daripada makna sebenarnya. Mereka memperkenalkan

counterfactual learning dan *entity deconfounding augmentation* untuk mengurangi ketergantungan model terhadap pola yang tidak relevan. Pendekatan ini terbukti membantu memperbaiki bias model, terutama pada teks berita yang memiliki struktur kalimat seragam.

Dalam konteks Bahasa Indonesia, ketidakseimbangan label juga menjadi masalah umum dalam pengembangan dataset NER. Budi dan Suryono (2023) menemukan bahwa dataset berbahasa Indonesia didominasi oleh entitas *Person* (PER) dan *Location* (LOC), sementara entitas seperti *Disease* (DIS) dan *Organization* (ORG) relatif sedikit. Ketimpangan ini membuat model sulit beradaptasi dengan domain baru seperti teks kesehatan. Temuan serupa dilaporkan oleh Yulianti dkk. (2024), yang menjelaskan bahwa dominasi kelas non-entitas dalam teks hukum menurunkan nilai *recall* pada entitas hukum. Penerapan *class weighting* saat pelatihan terbukti membantu meningkatkan performa model hingga 10% pada label minor.

Berbagai pendekatan telah dikembangkan untuk mengatasi masalah ini. Wang dkk. (2025b) memperkenalkan metrik *Entity Imbalance Degree* (EID) untuk mengukur tingkat ketimpangan label dalam dataset. Mereka menemukan bahwa dataset dengan nilai EID tinggi ketika label O mendominasi lebih dari 85% token cenderung menghasilkan model yang lemah dalam mendeteksi entitas minor meskipun menggunakan arsitektur yang kompleks. Belbekri dkk. (2024) kemudian mengembangkan metode *Two-Stage GAN Oversampling* yang memanfaatkan model GPT-3 dan *knowledge graph* DBpedia untuk membuat contoh sintetik yang menjaga kesesuaian semantik antartoken. Pendekatan ini membantu menyeimbangkan distribusi label tanpa merusak struktur bahasa.

Ketimpangan data juga menjadi isu umum di bahasa sumber rendah (*low-resource languages*) seperti Bahasa Indonesia. Haque dkk. (2021) melalui penelitian B-NER pada Bahasa Bangla menjelaskan bahwa sebagian besar dataset NER untuk bahasa lokal memiliki distribusi entitas yang tidak seimbang secara ekstrem. Mereka menekankan pentingnya pembangunan korpus beranotasi yang lebih besar dan beragam agar model tidak hanya belajar dari pola umum, tetapi juga mampu mengenali entitas langka dengan konteks yang lebih alami.

Secara keseluruhan, berbagai penelitian tersebut menunjukkan bahwa permasalahan ketidakseimbangan label dalam NER tidak dapat diatasi hanya dengan manipulasi

jumlah data. Diperlukan pendekatan yang memperhatikan keseimbangan antara proporsi data, konteks semantik, dan hubungan antarentitas agar model tidak hanya akurat pada kelas dominan, tetapi juga sensitif terhadap entitas minor yang memiliki nilai informatif tinggi. Pemahaman terhadap isu ini menjadi dasar penting dalam penelitian yang berupaya meningkatkan performa NER pada teks berita berbahasa Indonesia terutama ketika distribusi entitas tidak merata.

2.7 *Fine-Tuning* dan *Hyperparameter Optimization*

Fine-Tuning merupakan proses penyesuaian ulang model pra-latih (*pre-trained model*) agar mampu bekerja optimal pada domain atau dataset tertentu. Model pra-latih umumnya telah mempelajari pola bahasa umum melalui pelatihan skala besar, namun masih membutuhkan penyesuaian ketika diaplikasikan pada konteks yang berbeda. Proses *fine-tuning* bekerja dengan memperbarui parameter jaringan menggunakan sejumlah data tugas khusus, sambil mempertahankan pengetahuan dasar yang telah dipelajari sebelumnya. Menurut Benchama dan Zebbara (2024), *fine-tuning* dilakukan dengan hanya memperbarui sebagian lapisan atau keseluruhan parameter model, sehingga memungkinkan adaptasi lebih efisien tanpa harus melatih ulang dari awal yang memerlukan biaya komputasi lebih besar. Pendekatan ini membuat model lebih sensitif terhadap pola domain tertentu dan meningkatkan kemampuan generalisasi pada data nyata.

Kualitas *fine-tuning* sangat bergantung pada pengaturan *hyperparameter*, yaitu variabel pelatihan yang mengatur dinamika proses pembelajaran namun tidak dipelajari secara langsung oleh model. *Hyperparameter* mencakup komponen seperti *learning rate*, *batch size*, jumlah *epochs*, *dropout*, dan ukuran *hidden layer*. Masing-masing berperan menentukan dinamika pembaruan bobot selama pelatihan. Arai dkk. (2023) menekankan bahwa *learning rate* merupakan komponen paling krusial, dimana nilai terlalu besar membuat pembaruan bobot tidak stabil, sedangkan nilai terlalu kecil memperlambat proses dan berpotensi terjebak pada solusi suboptimal. Sementara itu, Ilemobayo dkk. (2024) menambahkan bahwa *batch size* berpengaruh terhadap stabilitas gradien, di mana *batch* besar menghasilkan gradien yang lebih stabil tetapi membutuhkan kapasitas komputasi tinggi, sedangkan *batch* kecil lebih cepat namun memiliki variasi gradien yang lebih besar. Selain itu, *dropout* digunakan untuk mencegah *overfitting* dengan menonaktifkan neuron secara acak selama pelatihan, dan jumlah *epoch* menentukan seberapa banyak

model meninjau ulang seluruh dataset. Evtimova dkk. (2023) menegaskan bahwa konfigurasi *hyperparameter* yang tepat sangat penting karena perubahan kecil pada komponen tersebut dapat menghasilkan perbedaan performa yang besar pada model.

Berbagai metode telah dikembangkan untuk menemukan konfigurasi *hyperparameter* terbaik. Metode paling tradisional adalah *grid search*, yaitu memeriksa semua kombinasi parameter dalam ruang pencarian. Meskipun sederhana, *grid search* sangat tidak efisien, terutama ketika ruang pencarian luas. Alternatif lainnya adalah *random search* yang memilih kombinasi parameter secara acak dari rentang nilai tertentu. *Random search* lebih efisien dibanding *grid search* karena memeriksa lebih banyak variasi parameter dalam waktu yang sama (Arai dkk., 2023). Meskipun demikian, kedua metode ini masih membutuhkan komputasi besar dan tidak memanfaatkan informasi performa dari percobaan sebelumnya.

Pendekatan terbaru banyak memanfaatkan metode optimasi berbasis pendekatan Bayesian, salah satunya *Tree-Structured Parzen Estimator* (TPE) yang diimplementasikan pada Optuna. Optuna merupakan *framework* otomatisasi pencarian hyperparameter yang menggabungkan efisiensi komputasi dan fleksibilitas pemodelan. Benchama dan Zebbara (2024) menunjukkan bahwa Optuna mampu memilih kombinasi hyperparameter optimal secara adaptif berdasarkan nilai objektif dari percobaan sebelumnya. TPE bekerja dengan memodelkan distribusi probabilitas dari nilai *hyperparameter* yang menghasilkan performa baik, lalu memprioritaskan area pencarian yang memiliki peluang menghasilkan peningkatan performa. Pendekatan ini membuat proses pencarian *hyperparameter* lebih efisien dibanding *random search* maupun *grid search*.

Selain keunggulan efisiensinya dalam proses pencarian parameter, Optuna juga menyediakan mekanisme *pruning*, yaitu penghentian percobaan lebih awal ketika performa model tidak menunjukkan peningkatan yang berarti. Fitur ini sangat bermanfaat dalam pelatihan model berukuran besar seperti Transformer atau BiLSTM yang membutuhkan waktu komputasi panjang. Kochnev dkk. (2025) menunjukkan bahwa integrasi Optuna dengan berbagai *framework deep learning* mampu menghemat waktu komputasi lebih dari 40% pada sejumlah eksperimen. Efektivitas tersebut semakin diperkuat oleh studi Kee dan Ho (2025), yang melaporkan bahwa Optuna dapat mempercepat proses *tuning* hingga 6 hingga 108 kali lebih cepat dibanding metode tradisional, sekaligus menghasilkan nilai *error* yang lebih rendah secara konsisten. Keunggulan Optuna juga terlihat pada

penelitian Benchama dan Zebbara (2024), yang mengintegrasikan Optuna dengan arsitektur CNN–BiGRU untuk sistem *intrusion detection*. Optuna digunakan untuk mengoptimalkan *learning rate*, jumlah *filter*, dan konfigurasi lain yang berdampak langsung pada peningkatan akurasi hingga mencapai 98,83% pada dataset NSLKDD. Temuan-temuan ini memperlihatkan bahwa Optuna tidak hanya efektif dalam menghasilkan kombinasi hyperparameter terbaik pada model *deep learning* yang kompleks untuk meningkatkan performa model, tetapi juga mampu memaksimalkan efisiensi penggunaan sumber daya komputasi.

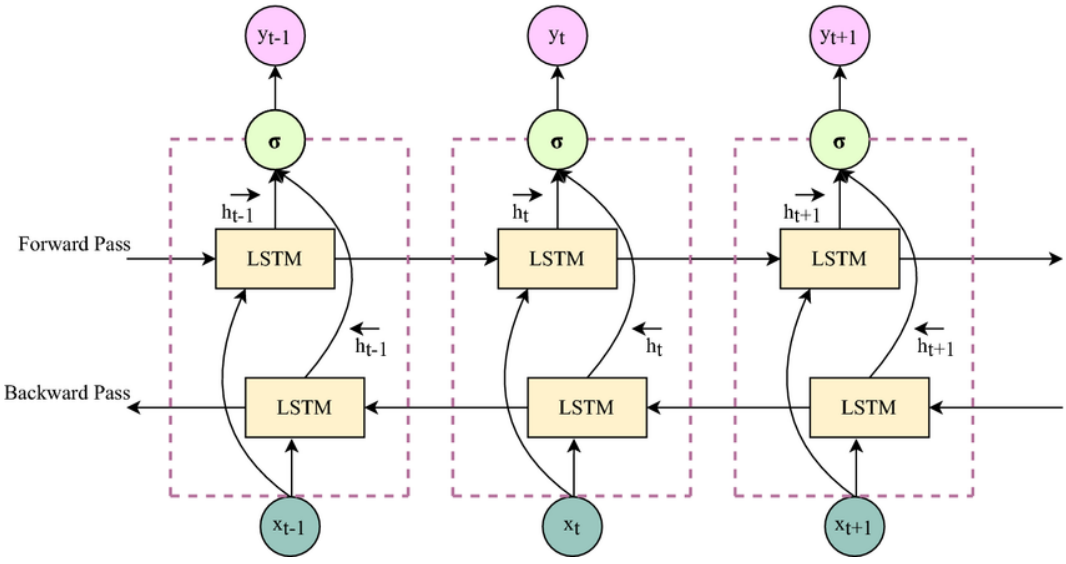
Dalam konteks penelitian NLP dan NER berbasis model hibrida seperti IndoBERT–BiLSTM, keberhasilan pelatihan sangat dipengaruhi oleh konfigurasi hyperparameter. Pemilihan *learning rate* yang tepat membantu model pra-latih beradaptasi secara stabil tanpa kehilangan representasi bahasa yang telah dipelajari sebelumnya. *Dropout* dan *batch size* membantu menjaga keseimbangan antara akurasi dan generalisasi, sedangkan jumlah *epoch* menentukan seberapa dalam model menyesuaikan diri dengan domain judul berita kesehatan. Oleh karena itu, penerapan *fine-tuning* yang tepat perlu dikombinasikan dengan optimasi hyperparameter yang sistematis agar model dapat mencapai performa optimal. *Framework* seperti Optuna mendukung proses ini melalui pencarian parameter secara efisien dan terarah, sehingga pelatihan dapat berlangsung lebih stabil, akurat, dan mudah direplikasi. Secara keseluruhan, *fine-tuning* dan *hyperparameter optimization* merupakan dua komponen fundamental dalam pengembangan model *deep learning* modern, terutama ketika diterapkan pada tugas NER berbahasa Indonesia yang membutuhkan adaptasi kontekstual yang kuat.

2.8 *Bidirectional Long Short-Term Memory* (BiLSTM)

Perkembangan *sequence learning* dalam pemrosesan bahasa alami bermula dari arsitektur RNN yang dirancang untuk menangani data berurutan seperti teks dan ucapan. RNN mampu memproses urutan pendek secara efektif, namun mengalami kesulitan dalam menangkap ketergantungan jangka panjang karena masalah *vanishing gradient*. Untuk mengatasi keterbatasan ini, Hochreiter dan Schmidhuber (1997) memperkenalkan *Long Short-Term Memory* (LSTM), yaitu varian RNN yang dilengkapi mekanisme *cell state* dan tiga gerbang utama' yaitu *input gate*, *forget gate*, dan *output gate* yang berfungsi mengatur aliran informasi serta menjaga konteks penting dalam urutan data. Arsitektur ini membuat LSTM

lebih stabil dan efektif ketika memproses teks panjang maupun tugas berbasis sekuens seperti *speech recognition* maupun *named entity recognition* (Gao dkk., 2021).

Arsitektur ini kemudian dikembangkan menjadi *Bidirectional* LSTM atau BiLSTM untuk memperluas kemampuan model dalam memahami konteks linguistik. BiLSTM memproses urutan kata dalam dua arah, yaitu *forward* (dari awal ke akhir) dan *backward* (dari akhir ke awal), sehingga mampu memanfaatkan informasi dari konteks sebelumnya dan sesudahnya secara bersamaan. Pendekatan dua arah ini membuat BiLSTM lebih efektif dalam memahami relasi antar-token dan makna kalimat secara keseluruhan (Shah dkk., 2022). Gao dkk. (2021) menegaskan bahwa BiLSTM memiliki kemampuan lebih kuat dalam menangkap informasi semantik pada kalimat panjang, karena setiap neuron mempertimbangkan konteks temporal dari dua arah secara simultan.



Gambar 2. Arsitektur *Bidirectional LSTM* (Naik & Jaidhar, 2022)

Gambar 2 menunjukkan arsitektur BiLSTM yang terdiri atas dua jalur pemrosesan, yaitu *forward pass* dan *backward pass*. Pada jalur *forward*, urutan data diproses dari kiri ke kanan ($x_{t-1} \rightarrow x_t \rightarrow x_{t+1}$), sedangkan pada jalur *backward*, data diproses dari kanan ke kiri. Kedua representasi ini kemudian digabungkan untuk menghasilkan keluaran dua arah (y_t), sehingga model dapat memahami informasi yang muncul sebelum dan sesudah token yang sedang diproses (Naik & Jaidhar, 2022).

Dalam ranah NLP, BiLSTM merupakan salah satu arsitektur populer sebelum hadirnya Transformer. Model ini sering digabungkan dengan komponen lain untuk meningkatkan performa. Shah dkk. (2022) mengembangkan BiLSTM–CNN untuk pengenalan entitas di domain e-commerce dan memperoleh akurasi 96,2% pada *Dark Web dataset* serta 92,9% pada CoNLL-2003, menunjukkan kemampuan BiLSTM menangkap pola semantik dan struktur teks semi-teratur seperti HTML. Pada domain medis, Deng dkk. (2021) menerapkan kombinasi BiLSTM–CRF pada teks pengobatan tradisional Tiongkok dan berhasil mendeteksi entitas penyakit dan gejala dengan akurasi lebih tinggi dibanding LSTM satu arah. Kolaborasi BiLSTM dengan model lain terus berkembang. Xu dan Li (2021) mengintegrasikan BERT dengan BiLSTM–CRF untuk ekstraksi entitas medis dan berhasil meningkatkan nilai *precision* dan *recall* dibanding BiLSTM–CRF konvensional. Zhou dkk. (2024) bahkan mengusulkan Multi-BiLSTM dengan *competition mechanism* untuk mengurangi *information loss* di setiap lapisan, yang menghasilkan prediksi lebih stabil pada domain medis.

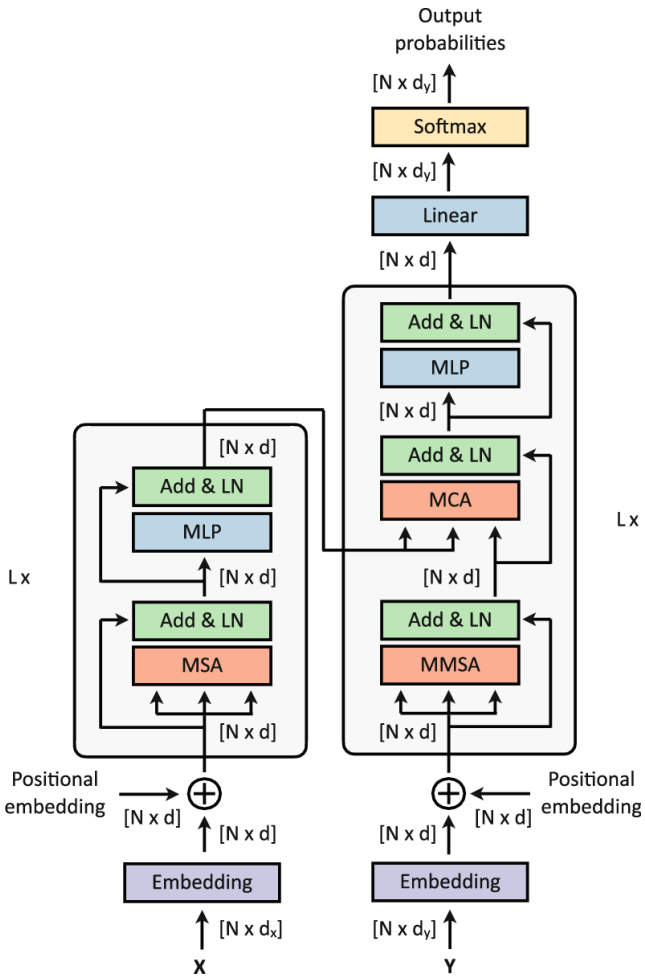
Dalam konteks Bahasa Indonesia, BiLSTM telah diterapkan untuk berbagai penelitian NER. Shidik dkk. (2024) mengembangkan model BiLSTM untuk mengenali entitas bencana dari korpus berita digital Indonesia, mencakup label baru seperti DISASTER, SCALE, SUPPLIES, dan CASUALTIES. Dengan teknik *random oversampling* untuk mengatasi ketidakseimbangan data, model mencapai *precision* 93,4%, *recall* 82,4%, dan F1-score 87,5%. Penelitian lain oleh Santoso dkk. (2021) menggunakan BiLSTM dengan pendekatan *end-to-end ontology* untuk mengekstraksi entitas dari berita daring dan menemukan bahwa model ini mampu mengatasi ambiguitas semantik yang sering muncul dalam teks berbahasa Indonesia. Hasil penelitian tersebut memperkuat pandangan bahwa BiLSTM memiliki fleksibilitas tinggi dan dapat digabungkan dengan model lain seperti CRF, CNN, maupun Transformer untuk meningkatkan ketepatan prediksi label antar-token.

2.9 *Transformer*

Model Transformer pertama kali diperkenalkan oleh Vaswani dkk. (2017) sebagai pendekatan baru dalam pemrosesan bahasa alami. Berbeda dari RNN dan LSTM yang memproses token secara berurutan, Transformer mengandalkan mekanisme *self-attention* yang memungkinkan model memahami hubungan antar-token secara

paralel. Pendekatan ini membuat Transformer mampu menangkap konteks jangka panjang tanpa batasan urutan linear, sehingga lebih cepat dan lebih akurat dalam memproses teks panjang.

Struktur dasar Transformer tersusun atas dua komponen utama, yaitu *encoder* dan *decoder*. Kedua komponen tersebut bekerja bersama untuk menghasilkan representasi kontekstual dan membentuk keluaran secara bertahap. *Encoder* tersusun dari lapisan *multi-head self-attention* (MSA) dan *feed-forward network* yang distabilkan oleh mekanisme *Add & Layer Normalization*. Sementara itu, *decoder* dilengkapi *masked multi-head self-attention* (MMSA) untuk memproses token keluaran secara autoregresif, serta *multi-head cross-attention* (MCA) yang menghubungkan informasi dari *encoder* (Courant dkk., 2023). Gambar 3 merupakan arsitektur dari transformer.



Gambar 3. *Transformer Architecture* (Courant dkk., 2023)

Transformer dipandang unggul berkat kemampuan paralelisasi dan *contextualized embedding*, yang merepresentasikan setiap token berdasarkan keseluruhan konteks kalimat, bukan hanya kata di sekitarnya (Wang dkk., 2024). Keunggulan tersebut menjadikan Transformer sebagai fondasi bagi berbagai model bahasa besar seperti BERT, GPT, dan RoBERTa (Gardazi dkk., 2025). Kemampuannya memahami makna dua arah (*bidirectional context*) dan mengolah data dalam skala besar menjadikan Transformer sebagai tonggak utama dalam perkembangan model pra-latih (*pre-trained language models*) modern.

2.10 *Bidirectional Encoder Representations from Transformers* (BERT)

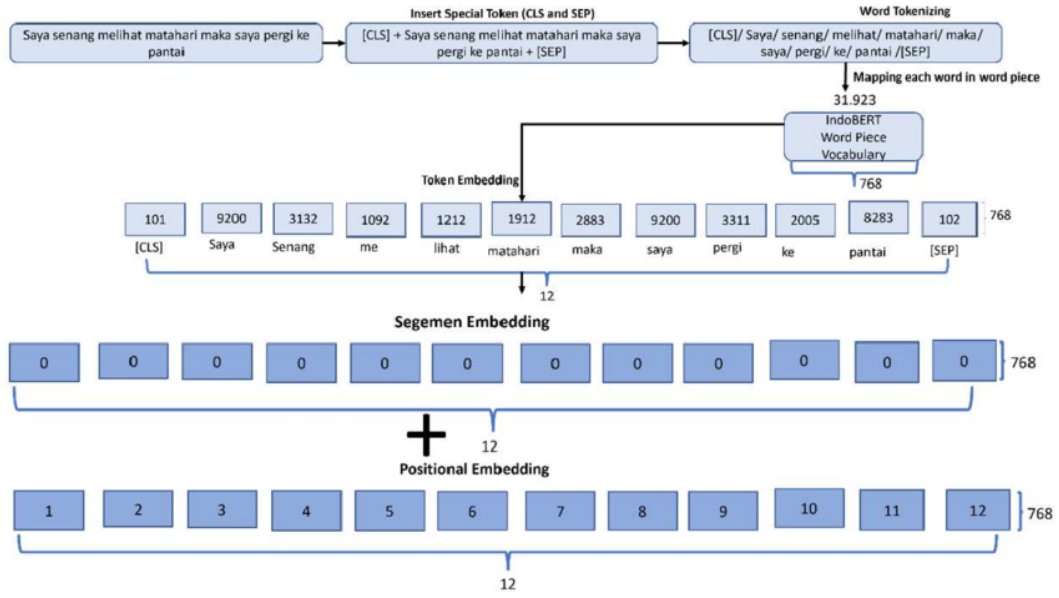
BERT dikembangkan oleh Devlin dkk. (2019) sebagai model berbasis *Transformer encoder* dua arah yang dirancang untuk memahami konteks kata dari sisi kiri dan kanan kalimat secara bersamaan. BERT dilatih melalui dua tugas utama, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Pada MLM, sebagian token disembunyikan (*masked*), kemudian model memprediksi token yang hilang berdasarkan konteks dua arah. Sedangkan NSP digunakan untuk mempelajari hubungan antarkalimat agar model mampu memproses struktur wacana secara lebih komprehensif.

Pendekatan tersebut menghasilkan representasi kata yang kontekstual dan adaptif, berbeda dengan model *embedding* statis seperti Word2Vec atau GloVe. Gardazi dkk. (2025) menyatakan bahwa *contextual embeddings* BERT memberikan hasil lebih akurat pada berbagai tugas NLP, termasuk NER, *Part-of-Speech Tagging*, dan *Sentiment Analysis*. Selain itu, model *encoder-based* seperti BERT memiliki keunggulan efisiensi dan keamanan dibandingkan model *decoder-only* seperti GPT, terutama untuk tugas ekstraksi informasi yang membutuhkan pemahaman konteks dua arah (Wang dkk., 2024).

Keunggulan lain BERT adalah sifat fleksibelnya dalam proses *fine-tuning*, sehingga dapat disesuaikan dengan berbagai bahasa dan domain. Dari model ini kemudian dikembangkan berbagai varian khusus seperti BioBERT, FinBERT, dan IndoBERT (Sayarizki dkk., 2024), yang disesuaikan dengan karakteristik korpus khusus sesuai bahasa maupun kebutuhan aplikasi tertentu.

2.10.1 IndoBERT

IndoBERT merupakan varian dari model BERT yang dikembangkan secara khusus untuk Bahasa Indonesia agar dapat memahami struktur morfologi, afiksasi, dan variasi gaya bahasa lokal secara lebih akurat. Model ini dikembangkan dalam dua keluarga besar, yaitu IndoBERT (IndoNLU) dan IndoBERT (IndoLEM). IndoBERT (IndoNLU) dilatih menggunakan korpus Indo4B yang berisi sekitar empat miliar kata (±23 GB, sekitar 250 juta kalimat) yang berasal dari berbagai sumber seperti berita daring, media sosial, Wikipedia, artikel daring, dan subtitle film, dengan beberapa varian seperti *base*, *large*, dan *lite* (berbasis ALBERT dengan *parameter-sharing*). Sementara itu, IndoBERT (IndoLEM) dilatih menggunakan sekitar 220 juta kata dari sumber seperti Wikipedia Indonesia, korpus berita (Kompas, Tempo, Liputan6), serta *Indonesian Web Corpus*. Pelatihan berbasis korpus monolingual ini memungkinkan IndoBERT menangkap pola linguistik khas Bahasa Indonesia secara lebih efektif dibanding model multibahasa seperti mBERT (Purnomo & Sutopo, 2024).



Gambar 4. Proses *Embedding IndoBERT* (Nabiilah dkk., 2024)

Gambar 4 memperlihatkan proses pembentukan *embedding* pada IndoBERT. Proses dimulai dengan penambahan token khusus [CLS] pada awal kalimat dan [SEP] pada akhir kalimat. Selanjutnya, kalimat diuraikan menjadi token menggunakan *WordPiece tokenizer*. Setiap token kemudian diubah menjadi *token embedding* berdimensi 768, yaitu ukuran standar yang digunakan pada arsitektur

BERT-base sebagai model dasar IndoBERT. Ukuran *embedding* ini bersifat desain arsitektur dan tidak diubah pada tahap *fine-tuning*, karena berkaitan langsung dengan jumlah parameter pada setiap lapisan Transformer. Meski demikian, ukuran *embedding* dapat berbeda pada varian model lain seperti BERT-large atau ALBERT yang menggunakan konfigurasi dimensi berbeda. Selain *token embedding*, *segment embedding* ditambahkan untuk membedakan token antarkalimat, sedangkan *positional embedding* menandai posisi setiap token dalam urutan. Ketiga *embedding* tersebut digabungkan membentuk *input embedding* yang selanjutnya diproses oleh lapisan-lapisan Transformer pada IndoBERT (Nabiilah dkk., 2024).

IndoBERT terbukti memberikan peningkatan performa pada tugas NER berbahasa Indonesia. Menurut Amien dan Gunawan (2024), *contextual embedding* IndoBERT lebih sesuai dengan struktur sintaksis dan semantik Bahasa Indonesia, sehingga meningkatkan akurasi ekstraksi entitas. Zainuddin dan Tahir (2025) juga melaporkan bahwa kombinasi IndoBERT dengan BiLSTM menghasilkan akurasi hingga 95,16% dan F1-score 94,51% pada ekstraksi entitas berita daring, jauh melebihi model konvensional seperti CRF dan BiLSTM tunggal. Selain itu, penelitian Budi dan Suryono (2023) dan Yulianti dkk. (2024) membuktikan bahwa IndoBERT unggul dalam berbagai domain, termasuk analisis berita, hukum, dan sosial, karena kemampuannya menghasilkan *contextual embedding* yang peka terhadap konteks, termasuk pada teks pendek seperti judul berita.

2.11 Model *Hybrid* NER

Model *hybrid* merupakan pendekatan dalam NLP yang menggabungkan dua atau lebih arsitektur pembelajaran untuk memanfaatkan keunggulan masing-masing model. Pendekatan ini berkembang karena satu jenis model saja sering kali tidak mampu memahami kompleksitas bahasa alami secara menyeluruh. Setiap arsitektur memiliki keunggulan dan keterbatasan masing-masing dalam memproses informasi linguistik. Transformer, misalnya, unggul dalam menangkap hubungan jarak jauh antar-token melalui mekanisme *self-attention*, sedangkan BiLSTM unggul dalam memahami pola urutan dua arah antarkata. Kombinasi kedua arsitektur tersebut menghasilkan representasi yang lebih komprehensif karena mampu menangkap konteks global sekaligus mempertahankan keterhubungan token dalam urutan kalimat (Wang dkk., 2023a; Ivanenko dkk., 2025).

Pada tugas *sequence labeling* seperti NER, pendekatan *hybrid* banyak digunakan karena masing-masing komponen memiliki fungsi yang saling melengkapi. Setiap arsitektur memberikan kontribusi berbeda dalam proses representasi teks. Model berbasis Transformer berperan menghasilkan representasi kontekstual dari teks melalui *contextual embeddings*, BiLSTM memproses urutan kata dari arah maju dan mundur untuk mempertahankan informasi posisi, dan CRF sering digunakan pada tahap akhir untuk memastikan urutan label yang konsisten dan sesuai struktur linguistik (Li dkk., 2024). Kombinasi ini telah terbukti meningkatkan stabilitas dan akurasi hasil pemodelan, karena setiap lapisan memperkuat pemahaman semantik dari sisi yang berbeda (Ivanenko dkk., 2025).

Kombinasi Transformer–BiLSTM sendiri dirancang untuk mengatasi dua permasalahan utama pada data teks, yaitu kebutuhan memahami konteks global sekaligus menjaga struktur lokal. Transformer melalui mekanisme *multi-head attention* dapat mengenali hubungan antar-token meskipun jaraknya jauh, sementara BiLSTM menjaga urutan konteks dua arah agar model tidak kehilangan hubungan semantik antarkata (Wang dkk., 2025a). Pendekatan ini menghasilkan representasi yang lebih stabil dan informatif dibandingkan model berurutan tradisional seperti RNN.

Berbagai penelitian menunjukkan efektivitas arsitektur *hybrid* ini. Struktur BERT–BiLSTM–CRF dilaporkan mampu meningkatkan nilai F1-score hingga 23% dibandingkan BiLSTM tunggal pada tugas pengenalan entitas medis (Gao dkk., 2021). Pendekatan serupa juga memberikan hasil lebih baik pada teks biomedis karena lapisan Transformer mampu menyajikan *embedding* dua arah yang kaya konteks sebelum diproses oleh BiLSTM (Xu & Li, 2021). Pendekatan Multi-BiLSTM dengan mekanisme kompetisi terbukti memperkuat interaksi antar-lapisan sehingga model dapat mempertahankan representasi sekuensial dengan lebih stabil (Zhou dkk., 2024). Selain itu, integrasi BERT–BiLSTM–CRF juga terbukti unggul pada tugas ekstraksi informasi di domain *e-commerce* dan teks multibahasa karena mampu memahami hubungan semantik yang kompleks dan menjaga struktur label tetap konsisten (Shah dkk., 2022; Lee dkk., 2022) .

Dalam Bahasa Indonesia, pendekatan *hybrid* semakin relevan karena karakteristik linguistik Bahasa Indonesia bersifat aglutinatif dan memiliki struktur kalimat yang fleksibel. Model seperti IndoBERT, yang dilatih menggunakan korpus besar Bahasa Indonesia, memberikan *contextual embeddings* yang sesuai dengan karakteristik

linguistik lokal (Amien & Gunawan, 2024). Ketika *embedding* tersebut dipadukan dengan BiLSTM, model memperoleh kelebihan tambahan berupa pemahaman hubungan antar-token yang lebih runtut dari dua arah.

Sejumlah penelitian menunjukkan bahwa *hybrid* IndoBERT–BiLSTM efektif pada berbagai tugas NLP berbahasa Indonesia. Kombinasi Transformer–BiLSTM pada ekstraksi entitas berita daring menghasilkan peningkatan akurasi yang signifikan (Zainuddin & Tahir, 2025). Arsitektur IndoBERT–BiLSTM memberikan performa kompetitif pada tugas klasifikasi teks Bahasa Indonesia (Setiawan dkk., 2024). Pada domain hukum, pendekatan IndoBERT–BiLSTM–CRF menghasilkan performa lebih baik dibandingkan model BiLSTM murni (Yulianti dkk., 2024). Selain itu, kombinasi Transformer dan BiLSTM dinilai efektif untuk dataset yang relatif kecil atau memiliki ketidakseimbangan label, karena masing-masing lapisan mampu saling mengompensasi kelemahan satu sama lain (Budi & Suryono, 2023). Secara keseluruhan, model *hybrid* menawarkan solusi yang kuat untuk tugas NER karena mampu memadukan pemahaman konteks global dari Transformer dan struktur sekuensial dari BiLSTM. Integrasi ini menjadikan pendekatan *hybrid* sebagai salah satu strategi yang paling banyak digunakan dalam pengembangan sistem ekstraksi informasi modern.

2.12 *Latent Dirichlet Allocation* (LDA)

LDA merupakan metode pemodelan topik berbasis probabilistik yang digunakan untuk menemukan struktur laten berupa kumpulan topik dalam dokumen teks yang tidak terstruktur. Setiap dokumen diasumsikan tersusun atas campuran beberapa topik, sedang kan setiap topik direpresentasikan sebagai distribusi probabilistik atas kata-kata.Pendekatan ini memungkinkan proses ekstraksi tema-tema inti dari kumpulan dokumen tanpa memerlukan anotasi manual, sehingga sangat bermanfaat dalam tugas eksplorasi data tekstual skala besar (Taher dkk., 2025).

LDA diperkenalkan sebagai model generatif probabilistik yang memodelkan bagaimana dokumen dihasilkan melalui proses acak berbasis topik laten. Dalam model ini, setiap dokumen d diasumsikan memiliki distribusi topik θ_d yang diambil dari distribusi Dirichlet dengan parameter α, sedangkan setiap topik k memiliki distribusi kata ϕ_k yang juga diambil dari distribusi Dirichlet dengan parameter β (Blei dkk., 2003). Proses generatif LDA secara matematis dirumuskan sebagai

berikut:

$$\theta_d \sim \text{Dirichlet}(\alpha), \qquad \phi_k \sim \text{Dirichlet}(\beta) \tag{1}$$

Untuk setiap token ke-n pada dokumen d, sebuah topik laten dipilih berdasarkan distribusi topik dokumen, yaitu:

$$z_{d,n} \sim \text{Multinomial}(\theta_d) \tag{2}$$

Selanjutnya, kata yang diamati dihasilkan berdasarkan distribusi kata dari topik terpilih tersebut, yang dirumuskan sebagai:

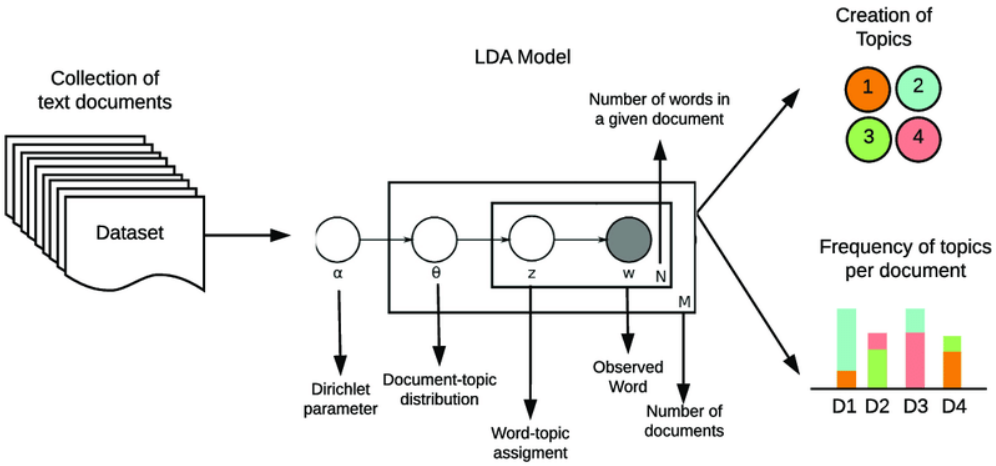
$$w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}}) \tag{3}$$

Dalam implementasinya, LDA bekerja berdasarkan asumsi *bag-of-words*, yaitu mengabaikan urutan kata dalam dokumen namun tetap mempertahankan frekuensi kemunculan kata. Dengan asumsi ini, setiap dokumen direpresentasikan sebagai vektor frekuensi kata:

$$\mathbf{x}_d = (x_{d,1}, x_{d,2}, \ldots, x_{d,V}) \tag{4}$$

dengan V menyatakan ukuran kosakata dan $x_{d,v}$ menunjukkan jumlah kemunculan kata ke-v dalam dokumen d. Representasi ini memungkinkan LDA mengidentifikasi pola ko-kemunculan kata yang konsisten dan memetakannya ke dalam topik-topik laten yang merepresentasikan tema semantik dokumen (Blei dkk., 2003).

Setelah proses inferensi dilakukan, model LDA menghasilkan dua keluaran utama, yaitu distribusi topik untuk setiap dokumen dan distribusi kata untuk setiap topik. Distribusi topik per dokumen menggambarkan proporsi keterlibatan masing-masing topik dalam dokumen, sedangkan distribusi kata per topik merepresentasikan kata-kata yang paling dominan dalam suatu topik. Karakter interpretatif ini memungkinkan setiap topik dipahami melalui kata-kata dengan probabilitas tinggi, sehingga LDA banyak digunakan dalam analisis konten, klasifikasi dokumen, dan peringkasan teks (Taher dkk., 2025). Gambar 5 menyajikan skematik algoritma LDA.



Gambar 5. Skematik Algoritma LDA (Buenano-Fernandez dkk., 2020)

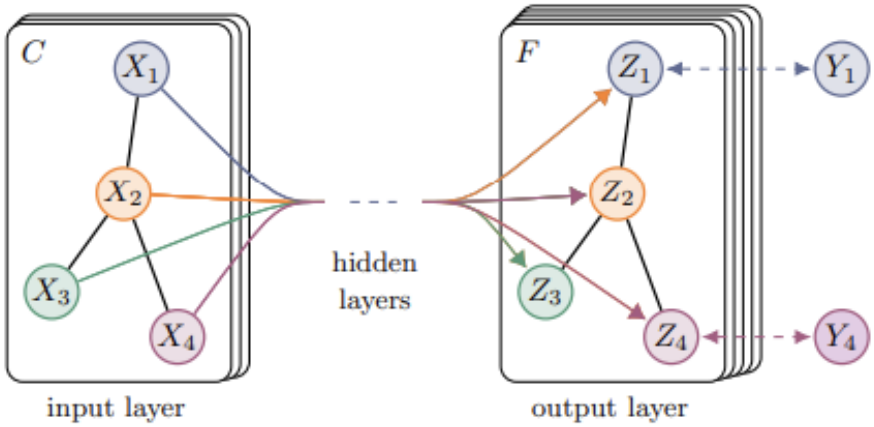
LDA dapat diterapkan pada berbagai domain. Dalam analisis berita daring, metode adaptif seperti *Adaptive LDA* digunakan untuk menentukan jumlah topik optimal sehingga struktur tematik dokumen lebih representatif (Batool dkk., 2024). Dalam penelitian yang berkaitan dengan pandemi COVID-19, LDA membantu mengungkap tema utama dalam percakapan publik dan isu lingkungan digital, seperti polusi perkotaan, kabut asap, dan deforestasi (Nuryana dkk., 2025). LDA juga dimanfaatkan untuk memetakan diskusi mengenai vaksin dan misinformasi kesehatan, sehingga mendukung pemantauan dinamika percakapan publik secara komputasional (Sharma dkk., 2023).

Selain untuk analisis topik, LDA dapat memperkuat teknik NLP lainnya. Dalam sistem ekstraksi entitas, LDA mampu menyediakan konteks semantik tambahan terutama pada korpus yang memiliki struktur lemah. LDA terbukti efektif membantu identifikasi entitas di domain pertanian (Gangadharan & Gupta, 2020) maupun dalam proses anotasi data *e-government*. Pada penelitian Umam dkk. (2025b), LDA digunakan sebagai *semantic pre-annotation tool* untuk membantu pelabelan NER. Distribusi topik memberikan konteks tambahan sehingga anotator lebih konsisten, terutama pada entitas yang jarang muncul. Integrasi ini meningkatkan kualitas anotasi dan performa model NER pada data publik. LDA juga dapat berperan dalam sistem penilaian konten, seperti pada proses *resume screening*, di mana representasi topik digunakan untuk menilai kecocokan kandidat berdasarkan keterampilan, pengalaman, atau bidang keahlian (Jagwani dkk., 2023). Pendekatan ini menunjukkan fleksibilitas LDA dalam menangani berbagai jenis teks dan mendukung alur kerja NLP di berbagai bidang.

Secara keseluruhan, LDA memiliki sejumlah keunggulan dalam analisis teks. Metode ini mampu mengekstraksi tema-tema tersembunyi dari dokumen tidak terstruktur sehingga mempermudah proses identifikasi pola dan pemetaan informasi dalam skala besar. Representasi topik yang dihasilkan juga dapat dimanfaatkan untuk menelusuri tren diskusi, mengelompokkan dokumen, serta mendukung berbagai tugas lanjutan dalam *text mining*. Selain itu, LDA dapat dipadukan dengan teknik NLP lainnya, termasuk NER, untuk memperkaya konteks semantik dan meningkatkan kualitas ekstraksi informasi. Fleksibilitas tersebut menjadikan LDA sebagai salah satu pendekatan dasar yang relevan dalam penelitian NLP modern, terutama dalam analisis judul berita daring, pemetaan isu publik, serta peningkatan akurasi ekstraksi entitas.

2.13 *Graph Convolutional Network* (GCN)

GCN merupakan arsitektur *Graph Neural Network* yang dirancang untuk memproses data berstruktur graf. Data berbasis graf memiliki relasi yang tidak beraturan dan bersifat kompleks, berbeda dari data sekuensial atau data berformat grid seperti teks atau citra. GCN memperbarui representasi setiap node dengan mengagregasi informasi dari node tetangga (*neighbor aggregation*), sehingga mampu menangkap pola relasional antarelemen dalam suatu graf. Pendekatan ini efektif dalam berbagai tugas NLP seperti NER, *relation extraction*, *sequence labeling*, dan *information extraction* karena mampu memodelkan konteks global serta dependensi jangka panjang yang tidak dapat ditangani secara optimal oleh model sekuens tradisional (Kipf & Welling, 2016). Gambar 6 menunjukkan arsitektur GCN.



Gambar 6. *Graph Convolutional Network* (Kipf & Welling, 2016)

Setiap node dalam graf memiliki fitur awal yang direpresentasikan sebagai vektor. Lapisan GCN kemudian melakukan propagasi informasi dengan mengagregasi fitur dari *node* dan tetangganya melalui operasi konvolusi berbasis graf. Proses ini menghasilkan representasi laten baru (misalnya Z_1, Z_2, Z_3, Z_4) yang lebih informatif karena telah mengintegrasikan struktur lokal dari graf. Dengan demikian, GCN mampu menangkap hubungan relasional antar-*node* secara efektif dan menghasilkan representasi yang lebih bermakna untuk tugas selanjutnya seperti klasifikasi atau ekstraksi informasi (Kipf & Welling, 2016).

Dalam pemrosesan bahasa alami, pendekatan berbasis graf digunakan untuk memodelkan hubungan sintaktik maupun semantik antartoken melalui *dependency tree* atau *semantic graph*. Setiap token direpresentasikan sebagai node, sedangkan relasi gramatikal seperti *nsubj*, *dobj*, dan *amod* direpresentasikan sebagai *edge*. Melalui agregasi informasi dari *node* tetangga, GCN mampu menangkap struktur dependensi yang kompleks dan memperkaya representasi token, terutama pada konteks yang tidak dapat ditangkap oleh model berbasis urutan atau *self-attention* saja (Zhang, 2021).

Salah satu tugas NLP yang sangat diuntungkan oleh pendekatan berbasis graf adalah NER karena banyak entitas dalam kalimat ditentukan oleh pola sintaktik dan bukan semata posisi token dalam teks. Chen dan Shen (2025) ditunjukkan bahwa integrasi fitur dependensi melalui GCN ke dalam arsitektur BiLSTM–CRF mampu meningkatkan kualitas representasi token, terutama untuk entitas seperti ORGANIZATION atau istilah teknis yang sangat bergantung pada struktur kalimat. Pengembangan lebih lanjut dilakukan pada Hanh dkk. (2021), yang menggabungkan *global embedding* berbasis GCN dengan *contextual embedding* dari XLNet. Kombinasi tersebut menghasilkan performa lebih tinggi pada dataset CoNLL-2003 dibandingkan penggunaan embedding statis atau kontekstual secara terpisah. Sementara itu, Zaratiana dkk. (2022) diperkenalkan GNNer, model *span-based NER* yang menggunakan GCN untuk memodelkan hubungan antarspan yang saling tumpang tindih. Pendekatan ini mengurangi prediksi entitas yang saling bertentangan melalui pemanfaatan graf overlap antarsapan.

Kinerja GCN dalam tugas *sequence labeling* juga ditingkatkan melalui pendekatan *nearest-example graph* seperti pada GNN-SL yang dikembangkan oleh Wang dkk. (2023b). Dalam pendekatan ini, GCN tidak hanya menghubungkan token dalam satu kalimat, tetapi juga menghubungkannya dengan contoh latih yang memiliki

pola serupa. Graf heterogen tersebut memungkinkan transfer informasi antarkalimat sehingga lebih efektif pada kondisi data terbatas, khususnya untuk entitas yang jarang muncul.

GCN juga sering dikombinasikan dengan arsitektur modern lain, termasuk Transformer. Integrasi GCN dengan Transformer memungkinkan model memperoleh representasi yang lebih stabil karena informasi sintaktik dari *dependency parsing* melengkapi keterbatasan Transformer dalam memodelkan hubungan struktural eksplisit (Zhang, 2021). GCN juga digunakan dalam tugas *multimodal information extraction*. Khanfir dkk. (2024) mengembangkan arsitektur yang memadukan *Sparse Graph Transformer Encoder* (SGTE) dengan GCN dalam *decoder* untuk memproses dokumen tulisan tangan. GCN membantu menyelaraskan fitur visual dan karakter sehingga pengenalan teks dan penandaan entitas menjadi lebih stabil. Pada penelitian lain, Zhou dkk. (2020) mengusulkan *Weighted GCN* dengan *Logical Adjacency Matrix* (LAM) untuk menangkap relasi multi-hop pada *relation extraction*. LAM memungkinkan *node* menggabungkan informasi dari node jauh dalam satu langkah propagasi sehingga mengatasi batasan GCN standar yang hanya memproses ketetanggaan tingkat pertama.

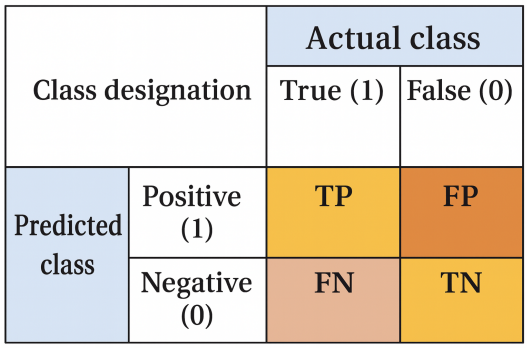
Secara keseluruhan, GCN berperan penting dalam pengolahan bahasa alami modern. Pendekatan ini efektif dalam menangkap struktur sintaksis yang tidak dapat ditangkap oleh model sekuensial, memodelkan hubungan global antartoken maupun antarspan, serta memperkuat representasi pada tugas multimodal. Integrasi GCN dengan arsitektur seperti BiLSTM, CRF, Transformer, dan XLNet menunjukkan bahwa pemodelan berbasis graf merupakan komponen penting dalam NLP modern, terutama untuk tugas yang menuntut pemahaman relasional yang lebih dalam.

2.14 Evaluasi Model

Evaluasi model merupakan tahapan penting untuk menilai sejauh mana sistem klasifikasi mampu menghasilkan prediksi yang mendekati label sebenarnya. Pada tugas pemrosesan bahasa alami seperti NER, proses evaluasi dilakukan dengan mengukur ketepatan model dalam mengidentifikasi entitas di dalam teks. Komponen dasar evaluasi terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Keempat komponen tersebut menjadi dasar dalam menghitung berbagai metrik performa yang digunakan secara luas pada penelitian

klasifikasi modern (Vujović, 2021).

Penilaian performa model umumnya menggunakan confusion matrix seperti pada Gambar 7



Gambar 7. *Confusion Matrix* (Vujović, 2021)

Melalui struktur matriks tersebut, empat komponen evaluasi didefinisikan sebagai berikut:

- *True Positive* (TP): model memprediksi kelas positif dan label sebenarnya memang positif.
- *False Positive* (FP): model memprediksi positif, tetapi label sebenarnya negatif.
- *False Negative* (FN): model memprediksi negatif, tetapi label sebenarnya positif.
- *True Negative* (TN): model memprediksi negatif dan label sebenarnya juga negatif.

Definisi ini menjadi landasan perhitungan berbagai metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score* (Hicks dkk., 2022).

Metrik paling dasar adalah *accuracy*, yaitu proporsi prediksi benar dibandingkan seluruh data. Rumusnya dituliskan sebagai berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Meskipun mudah dipahami, *accuracy* sering kali menyesatkan jika digunakan pada data yang tidak seimbang. Pada kondisi tersebut, kelas dengan jumlah data paling banyak dapat mendominasi hasil sehingga tampak seolah-olah model memiliki performa tinggi, padahal sebenarnya hanya mengikuti kelas mayoritas (Sathyanarayanan & Tantri, 2024).

Penilaian performa yang lebih informatif dilakukan melalui *precision* dan *recall*,

terutama ketika distribusi label tidak merata. *precision* mengukur proporsi prediksi positif yang benar, dirumuskan sebagai berikut:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Nilai *Precision* yang tinggi menandakan bahwa model jarang melakukan kesalahan dengan memprediksi positif pada sampel yang sebenarnya negatif (*false positive*). Sebaliknya, *recall* mengukur kemampuan model dalam menemukan seluruh sampel yang benar-benar positif dan dirumuskan sebagai berikut:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Tingkat *Recall* yang tinggi menunjukkan bahwa model mampu meminimalkan kesalahan *false negative*. Karena *precision* dan *recall* sama-sama penting, keduanya digabungkan dalam metrik *F1-score*, yaitu harmonisasi antara keduanya:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Menurut Sathyanarayanan dan Tantri (2024), *F1-score* merupakan metrik yang paling representatif untuk data dunia nyata yang umumnya tidak seimbang, sehingga lebih tepat dibandingkan hanya mengandalkan *accuracy* saja.

Selain sebagai dasar perhitungan metrik, *confusion matrix* juga memberikan visualisasi pola kesalahan model. Melalui struktur matriks, peneliti dapat melihat apakah model cenderung bias terhadap kelas tertentu ataupun ketidakmampuan model dalam menangani kelas minoritas. Kemampuan ini menjadikan *confusion matrix* alat analisis performa yang sangat informatif dalam sistem klasifikasi (Hicks dkk., 2022).

Pada kasus *multi-label classification*, *confusion matrix* diperluas untuk menghitung TP, FP, FN, dan TN untuk setiap label secara terpisah. Heydarian dkk. (2022) menekankan bahwa pendekatan *multi-label confusion matrix* penting untuk aplikasi seperti klasifikasi teks, kesehatan, dan multi-label NER yang melibatkan lebih dari satu label per sampel.

Dataset dengan distribusi label yang tidak seimbang memerlukan metrik evaluasi yang sensitif terhadap kelas minoritas. Dalam kondisi seperti ini, *F1-score* menjadi metrik paling sesuai karena memperhitungkan keseimbangan antara *precision* dan *recall* serta mampu menangkap kesalahan pada entitas yang jarang muncul (Sathyanarayanan & Tantri, 2024). Dengan demikian, kombinasi *confusion matrix*, *precision*, *recall*, dan *F1-score* memberikan kerangka evaluasi yang komprehensif dan akurat untuk menilai performa model klasifikasi pada teks, termasuk tugas NER.

BAB III

METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Waktu dan Tempat Penelitian ini yaitu sebagai berikut:

3.1.1 Tempat Penelitian

Penelitian ini dilaksanakan secara studi literatur di jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung. Lokasi bertempat di Jalan Prof. Dr. Soemantri Brojonegoro No.1, Gedong Meneng, Bandar Lampung.

3.1.2 Waktu Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun ajaran 2024/2025 hingga semester genap tahun ajaran 2025/2026, dengan rentang waktu pelaksanaan dari Agustus 2024 sampai Februari 2026. Kegiatan penelitian dibagi ke dalam tiga tahap utama.

Tahap pertama meliputi penentuan topik, studi literatur, pengumpulan referensi, eksplorasi data, serta penyusunan Bab I–III. Tahap kedua mencakup pengolahan data, pembangunan dan pelatihan model Hybrid IndoBERT–BiLSTM, serta evaluasi performa model. Tahap ketiga meliputi penyusunan Bab IV dan V, seminar hasil, dan persiapan sidang komprehensif.

3.2 Data dan Alat

3.2.1 Data

Penelitian ini memanfaatkan dataset InaCOVED (*Indonesian COVID-19 Online News and Entities Dataset*), yaitu kumpulan judul berita daring berbahasa Indonesia yang berfokus pada pemberitaan penyakit menular, khususnya COVID-19. Dataset ini dikembangkan oleh Badan Riset dan Inovasi Nasional (BRIN) melalui Pusat Riset Sains Data dan Informasi KST Samaun Samadikun. Akses terhadap dataset bersifat internal dan tidak tersedia secara terbuka, meskipun sumber beritanya berasal dari media daring nasional.

Data dikumpulkan dalam rentang waktu Januari hingga Mei 2020, yaitu periode awal pandemi COVID-19 di Indonesia ketika pemberitaan mengenai isu kesehatan meningkat secara luas di berbagai media nasional. Sumber berita mencakup berbagai portal berita daring nasional yang telah berperan dalam penyebaran informasi terkait pandemi. Keberagaman media tersebut memberikan variasi dalam gaya penulisan, struktur kalimat, serta cara penyajian informasi kesehatan di ruang publik. Informasi yang termuat dalam judul berita meliputi perkembangan jumlah kasus, wilayah terdampak, kebijakan penanganan pemerintah, kerja sama antarnegara, serta peran lembaga kesehatan dalam merespons pandemi.

Jumlah keseluruhan data yang digunakan adalah 16.839 judul berita. Penelitian ini secara khusus memusatkan analisis pada judul karena sifatnya yang singkat dan padat informasi. Keterbatasan konteks pada judul menjadikan proses pengenalan entitas lebih menantang dibandingkan teks berita penuh. Secara statistik, panjang judul rata-rata terdiri atas 10,22 token dengan median 10 token, panjang minimum 3 token, dan maksimum 27 token. Rentang tersebut menunjukkan variasi kompleksitas struktur dalam teks yang relatif pendek.

Struktur awal dataset memuat atribut *title_clean* dan *portal*. Atribut *title_clean* berisi judul yang telah melalui tahap pembersihan teks, sedangkan atribut *portal* menunjukkan asal media daring. Untuk keperluan tugas *Named Entity Recognition* (NER), data kemudian diformat mengikuti standar CoNLL dengan pelabelan entitas pada tingkat token menggunakan skema BIO (*Beginning, Inside, Outside*).

Proses pelabelan dilakukan secara manual dengan bantuan perangkat *Label Studio* dan melibatkan empat anotator. Setiap anotator bekerja berdasarkan pedoman anotasi yang telah ditetapkan guna menjaga konsistensi interpretasi entitas. Kategori entitas yang digunakan meliputi empat kelas utama yang relevan dengan konteks penyakit menular, yaitu PER (*Person*), ORG (*Organization*), LOC (*Location*), dan DIS (*Disease*).

Konsistensi anotasi diuji melalui pengukuran reliabilitas antar anotator menggunakan koefisien Cohen's Kappa. Perhitungan menghasilkan nilai 0,9567 dengan *observed agreement* sebesar 0,9818. Berdasarkan kriteria interpretasi Landis dan Koch, nilai tersebut termasuk dalam kategori *almost perfect agreement*, yang menunjukkan tingkat kesepakatan yang sangat tinggi. Hasil ini mengindikasikan bahwa proses anotasi dilakukan secara stabil dan dapat dipercaya sebagai dasar pelatihan serta evaluasi model.

3.2.2 Alat

3.2.2.1 Spesifikasi Perangkat

Penelitian ini menggunakan sebuah laptop ASUS dengan spesifikasi perangkat sebagai berikut:

- Merek / Model : ASUS VivoBook X513EAN
- Pabrikan : ASUSTeK COMPUTER INC.
- Prosesor : 11th Gen Intel® Core™ i5-1135G7 @ 2.40 GHz
- Jumlah Core : 4 core, 8 threads
- Memori (RAM) : 8 GB
- Sistem Operasi : Windows 11 Home Single Language 64-bit (Build 26200)
- DirectX Version : DirectX 12
- BIOS : X513EAN.301

Spesifikasi ini digunakan untuk menjalankan proses pemrograman, pemrosesan data, pelatihan model, serta evaluasi hasil penelitian.

3.2.2.2 Software Penelitian

Software yang digunakan dalam penelitian ini meliputi:

- Sistem Operasi : Windows 11
- Bahasa Pemrograman : Python versi 3.11.12
- Platform Eksekusi : Google Colab (untuk proses komputasi berbasis GPU/TPU)

Beberapa library Python yang digunakan dalam penelitian ini ditampilkan pada Tabel 2.

Tabel 2. Library Python yang Digunakan dalam Penelitian

No	Library	Versi	Fungsi
1	matplotlib	3.9.2	Visualisasi grafik seperti kurva *loss*, F1-score, dan *confusion matrix*.
2	numpy	1.26.4	Operasi numerik, manajemen *array*, dan normalisasi matriks *adjacency* pada GCN.
3	optuna	3.6.1	Optimasi *hyperparameter* seperti *learning rate*, *dropout*, *batch size*, dan jumlah unit LSTM.
4	pandas	2.2.2	Pengelolaan dataset CoNLL, pemrosesan data tabular, serta ekspor/impor data dalam format CSV.
5	scikit-learn	1.6.1 / 1.5.1	Pembagian data, perhitungan metrik evaluasi, *CountVectorizer*, dan *LatentDirichletAllocation* untuk LDA.
6	seaborn	0.13.2	Visualisasi *heatmap confusion matrix* untuk evaluasi performa model.
7	seqeval	1.2.2	Perhitungan metrik *sequence labeling* berbasis skema BIO.
8	stanza	1.8.2	*Dependency parsing* Bahasa Indonesia untuk membangun *dependency graph* pada arsitektur GCN.
9	torch	2.4.1 / 2.3.1	Framework *deep learning* untuk pelatihan model NER (IndoBERT–BiLSTM/GCN), operasi tensor, dan dukungan GPU.
10	transformers	4.44.2 / 4.43.4	Tokenisasi WordPiece IndoBERT serta menghasilkan *embedding* dari model pra-latih.

3.3 Metode Penelitian

Secara umum, penelitian ini dilakukan melalui beberapa tahapan sebagai berikut.

1. Input Data
   Data yang digunakan adalah dataset InaCOVED, berisi 16.839 judul berita daring berbahasa Indonesia dengan anotasi BIO dalam format CoNLL-2003. Dataset diperoleh dari BRIN sehingga proses anotasi tidak perlu dilakukan ulang.

2. *Pre-processing* Data
   Tahapan ini dilakukan untuk memastikan kualitas dan kebersihan data. Langkah-langkah yang dilakukan:

- Menghapus judul duplikat.
- Normalisasi data

3. Konversi Format Label: BIO $\rightarrow$ BIOES
   Konversi anotasi dilakukan untuk mengubah format BIO menjadi BIOES dengan menerapkan aturan transformasi sebagai berikut:

   - B- (*Begin*)
     Diubah menjadi S- (*Single*) apabila token tersebut merupakan entitas tunggal yang tidak diikuti oleh I- dengan tipe entitas yang sama.
     Tetap B- pada format BIOES apabila token tersebut merupakan awal rangkaian entitas multi-token dan diikuti oleh I-.
   - I- (*Inside*)
     Tetap I- apabila masih berada di tengah rangkaian entitas dan diikuti oleh I- lain.
     Diubah menjadi E- (*End*) apabila merupakan token terakhir dari rangkaian entitas dan diikuti oleh O atau label berbeda
   - O (*Outside*)
     Tetap O karena label ini tidak mengalami perubahan pada format BIOES.

4. Pemeriksaan distribusi label
   Setelah proses konversi, dilakukan pengecekan distribusi label pada dataset untuk mengidentifikasi ketidakseimbangan label. Proses ini bertujuan untuk mengevaluasi sebaran label entitas, baik pada format BIO, BIOES, maupun Entity, guna mengetahui sejauh mana ketidakseimbangan distribusi label yang ada dalam dataset.

5. Split Data
   Pembagian dataset dilakukan secara *stratified* agar proporsi setiap label tetap terjaga. Adapun komposisinya adalah sebagai berikut:

   - 80% data model
   - 20% data *testing*

   Selanjutnya 80% data model dibagi menjadi:

   - 80% *training*
   - 20% *validation*

6. Penanganan Ketidakseimbangan Label
   LDA digunakan untuk mengatasi ketidakseimbangan label dengan menambahkan distribusi topik pada setiap judul. Informasi topik tersebut

memperkaya representasi semantik sehingga token dengan label jarang dan token O yang berada di sekitar entitas memperoleh konteks tambahan yang membuatnya lebih informatif.

7. Pemeriksaan Ulang Distribusi Label Setelah LDA
   Tahapan ini dilakukan untuk mengevaluasi dampak penerapan LDA terhadap distribusi representasi token, khususnya token O yang sebelumnya mendominasi dataset.

8. Tokenisasi
   Tokenisasi dilakukan menggunakan IndoBERT tokenizer dengan langkah:
   - *WordPiece tokenization*
   - Penambahan token [CLS] dan [SEP]
   - *Truncation* untuk membatasi panjang maksimal sekuens

9. *Padding*
   Padding diterapkan untuk menyamakan panjang seluruh sekuens dengan menambahkan token pad hingga mencapai panjang maksimum yang ditentukan.

10. *Embedding*
    Embedding IndoBERT menghasilkan representasi kontekstual berdimensi 768 sebagai masukan awal model.

11. Pendefinisian Model
    Model yang dibangun terdiri dari beberapa arsitektur, yaitu:
    - IndoBERT-BILSTM
    - LDA +(IndoBERT-BILSTM)
    - (IndoBERT-BILSTM) + GCN
    - LDA + (IndoBERT-BILSTM) + GCN

12. *Hyperparameter Tuning* (Optuna)
    Optimasi dilakukan menggunakan Optuna dengan konfigurasi sebagai berikut:
    - Total 100 *trial*
    - Parameter yang dioptimasi:
      – *learning rate*
      – *dropout*
      – *batch size*
      – jumlah unit LSTM

- jumlah *layer* LSTM
- jumlah unit *dense layer*

Tujuan optimasi adalah mendapatkan kombinasi parameter yang memberikan performa terbaik pada data validasi.

13. Pelatihan dan Evaluasi Model
    Pelatihan model dilakukan menggunakan model yang sudah didefinisikan pada tahapan sebelumnya, dengan parameter hasil dari tahapan *hyperparameter tuning*.

    Pengaturan pelatihan:

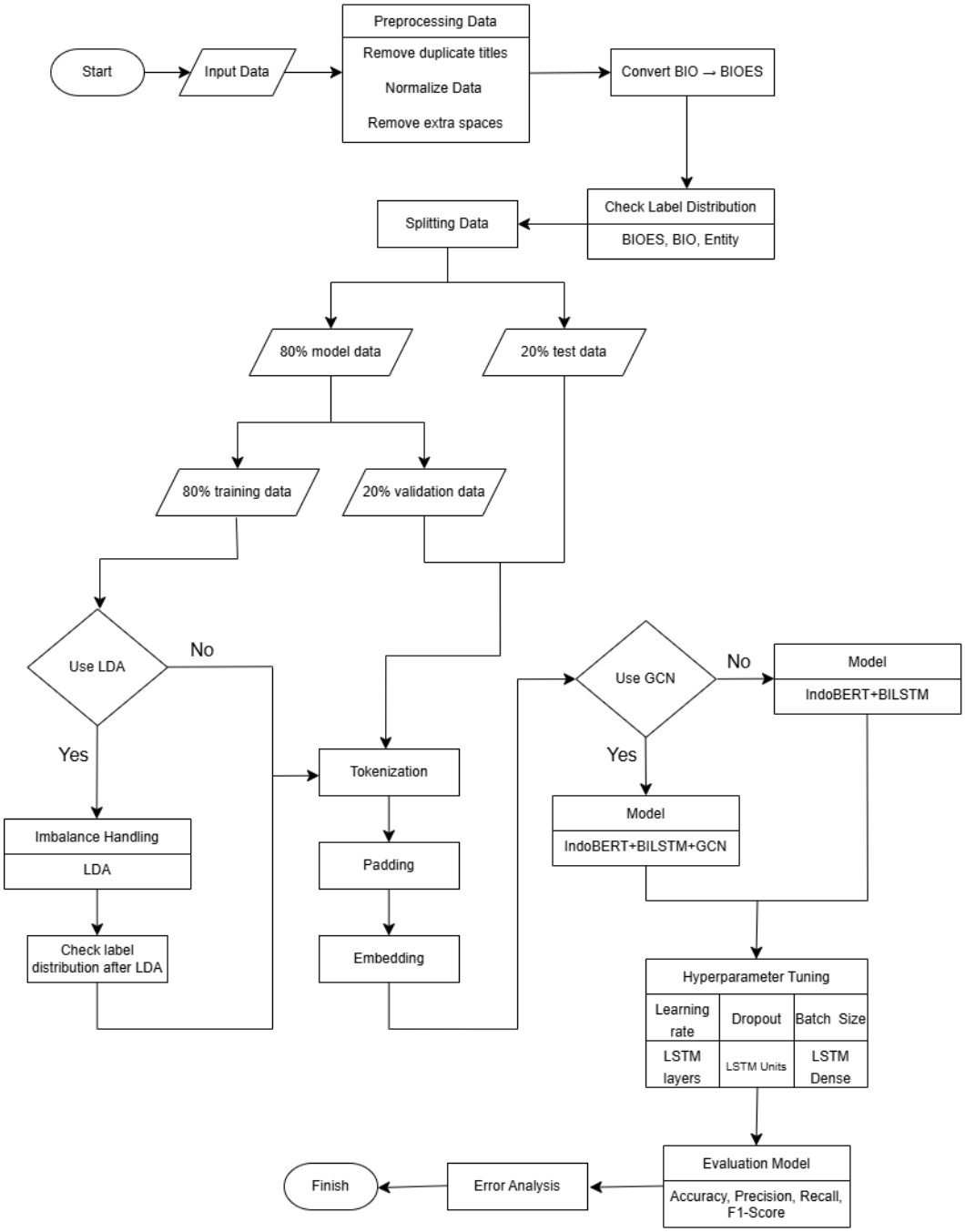
    - Maksimum 50 *epoch*
    - Menggunakan *Early Stopping* berdasarkan *validation macro F1-score*
    - Optimizer: AdamW
    - Loss function: *Focal CrossEntropyLoss*

    Evaluasi dilakukan menggunakan metrik:

    - Accuracy
    - Precision
    - Recall
    - F1- score
    - Visualisasi *confusion matrix*

14. *Error* Analisis
    Analisis kesalahan dilakukan untuk mengidentifikasi pola kegagalan model serta menilai kontribusi LDA dan GCN terhadap perbaikan performa. Tahapan ini mencakup pemeriksaan kesalahan prediksi pada tingkat token.

Analisis kontribusi setiap komponen dalam sistem yang diusulkan dilakukan melalui empat skenario eksperimen, yaitu model *baseline* IndoBERT–BiLSTM, IndoBERT–BiLSTM dengan LDA, IndoBERT–BiLSTM dengan GCN, serta kombinasi IndoBERT–BiLSTM, LDA, dan GCN. Seluruh skenario menggunakan tahapan pemrosesan dan konfigurasi evaluasi yang sama, sehingga perbedaan performa mencerminkan pengaruh masing-masing metode. Kerangka penelitian untuk keempat skenario tersebut sebagaimana ditunjukkan pada Gambar 8 berikut:



Gambar 8. Alur Penelitian

Berdasarkan alur penelitian di atas, penelitian ini menguji empat skenario eksperimen sebagai berikut:

- **LDA = No, GCN = No:** IndoBERT + BiLSTM (*baseline*)
- **LDA = Yes, GCN = No:** LDA + IndoBERT + BiLSTM
- **LDA = No, GCN = Yes:** IndoBERT + BiLSTM + GCN

- **LDA = Yes, GCN = Yes:** LDA + IndoBERT + BiLSTM + GCN

Setiap skenario menguji kombinasi teknik yang berbeda untuk meningkatkan performa model dalam tugas NER.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini mengkaji penerapan *Named Entity Recognition* (NER) pada judul berita daring berbahasa Indonesia di domain kesehatan dengan mempertimbangkan keterbatasan konteks teks serta ketidakseimbangan distribusi label. Pendekatan yang digunakan melibatkan model hibrida IndoBERT–BiLSTM sebagai *baseline*, serta pengayaan semantik dan struktural melalui integrasi *Latent Dirichlet Allocation* (LDA) dan *Graph Convolutional Network* (GCN). Berdasarkan hasil eksperimen dan analisis yang telah dipaparkan pada Bab IV, diperoleh beberapa kesimpulan utama sebagai berikut.

Pertama, penerapan model hibrida IndoBERT–BiLSTM terbukti mampu melakukan NER secara efektif pada judul berita daring yang bersifat singkat dan padat. Kombinasi representasi kontekstual dari IndoBERT dan pemodelan urutan token melalui BiLSTM menghasilkan prediksi yang stabil dalam mengenali tipe entitas secara global. Pada evaluasi menggunakan skema entitas (PER, LOC, ORG, DIS, dan O), model *baseline* menunjukkan kinerja yang paling konsisten dengan nilai *macro F1-score* tertinggi dibandingkan model lainnya. Temuan ini mengindikasikan bahwa pada teks dengan keterbatasan konteks, pemanfaatan *contextual embedding* yang kuat berperan penting dalam menjaga ketepatan prediksi serta menekan kesalahan pada token non-entitas. Namun demikian, hasil ini juga menunjukkan kecenderungan model *baseline* untuk lebih konservatif dan bias terhadap kelas dominan O, sehingga sensitivitas terhadap entitas dengan frekuensi rendah masih terbatas.

Kedua, integrasi LDA dan GCN memberikan kontribusi yang berbeda dalam menangani ketidakseimbangan label serta memperkaya representasi token. Penerapan LDA meningkatkan sensitivitas model terhadap keberadaan entitas, khususnya pada kelas Disease (DIS), Location (LOC), dan Person (PER). Informasi

topik global dari LDA memberikan sinyal semantik tambahan pada token yang berada dalam konteks kesehatan, sehingga model menjadi lebih jarang melewatkan entitas yang seharusnya dikenali, termasuk entitas yang relatif jarang muncul pada data. Kondisi ini tercermin dari meningkatnya nilai *recall* pada hampir seluruh kelas entitas. Meskipun demikian, peningkatan sensitivitas tersebut juga diiringi oleh menurunnya selektivitas, di mana sebagian token non-entitas ikut diprediksi sebagai entitas. Hal ini menunjukkan bahwa pengayaan semantik berbasis topik cenderung menggeneralisasi konteks secara luas dan memerlukan mekanisme pengendalian agar tidak memicu *over-detection*.

Sebaliknya, integrasi GCN berkontribusi dalam memperbaiki pemisahan tipe entitas dan menjaga konsistensi prediksi melalui pemodelan relasi struktural antartoken. Model dengan GCN menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall* dibandingkan model yang hanya mengandalkan LDA. Pemanfaatan struktur graf membantu model membedakan entitas dengan kemiripan semantik, seperti PER dan ORG, berdasarkan posisi serta hubungan token dalam kalimat, bukan semata-mata berdasarkan konteks tematik global. Dengan demikian, GCN mampu meningkatkan sensitivitas terhadap entitas minor tanpa menyebabkan lonjakan kesalahan prediksi yang berlebihan. Meskipun demikian, tantangan masih ditemukan pada konsistensi token internal entitas, yang mengindikasikan perlunya mekanisme pengendalian *boundary* yang lebih eksplisit.

Ketiga, perbandingan kinerja antar model menunjukkan bahwa peningkatan nilai metrik evaluasi secara numerik tidak selalu mencerminkan kualitas model secara menyeluruh. Model *baseline* IndoBERT–BiLSTM menunjukkan performa paling stabil dari sisi akurasi global dan ketepatan prediksi terhadap token non-entitas. Namun, model ini cenderung lebih konservatif dan memiliki sensitivitas yang lebih rendah terhadap entitas dengan frekuensi kemunculan yang kecil.

Model berbasis LDA dan model yang mengintegrasikan LDA dan GCN menunjukkan sensitivitas yang lebih tinggi terhadap entitas, terutama pada kelas-kelas minor. Hal ini tercermin dari meningkatnya nilai *recall*, yang berarti model lebih jarang melewatkan entitas yang seharusnya dikenali. Dalam konteks pemantauan isu kesehatan, kemampuan ini memiliki nilai praktis karena dapat mengurangi kehilangan informasi penting. Model yang menggabungkan LDA dan GCN secara khusus menunjukkan cakupan deteksi entitas yang paling luas, karena sinyal semantik global dan relasi struktural antartoken bekerja secara simultan dalam

mengaktifkan kandidat entitas.

Meskipun demikian, peningkatan sensitivitas pada model berbasis LDA dan model gabungan diiringi oleh peningkatan kesalahan prediksi pada token non-entitas, yang menunjukkan adanya *trade-off* antara cakupan deteksi dan ketepatan klasifikasi. Sementara itu, model IndoBERT–BiLSTM dengan integrasi GCN menunjukkan keseimbangan terbaik antara peningkatan sensitivitas terhadap entitas minor dan pengendalian kesalahan prediksi. Model ini mampu memperluas deteksi entitas dibandingkan *baseline*, namun tetap menjaga konsistensi dan stabilitas prediksi secara keseluruhan.

Dengan demikian, apabila tujuan sistem adalah memaksimalkan cakupan deteksi entitas (*high-recall scenario*), maka model yang mengintegrasikan LDA dan GCN menjadi pilihan yang relevan. Namun, dalam konteks penelitian ini yang menekankan penanganan ketidakseimbangan label tanpa mengorbankan stabilitas prediksi, model dengan integrasi GCN dapat dipandang sebagai pendekatan yang paling selaras dan representatif.

5.2 Saran

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, terdapat beberapa saran yang dapat dipertimbangkan untuk pengembangan penelitian dan penerapan sistem NER pada masa mendatang.

Pertama, penelitian selanjutnya disarankan untuk menambahkan mekanisme pengendalian batas entitas (*boundary*) yang lebih eksplisit, seperti *Conditional Random Field* (CRF) atau pendekatan berbasis *span-level classification*. Integrasi mekanisme tersebut berpotensi memperkuat konsistensi struktur entitas multi-token, khususnya pada token internal (*I-tag*) yang masih menjadi sumber kesalahan dominan pada skema BIOES. Dengan adanya pengendalian batas yang lebih terstruktur, peningkatan sensitivitas deteksi dapat tetap dipertahankan tanpa mengorbankan stabilitas struktur entitas.

Kedua, pemanfaatan LDA sebagai pengaya konteks semantik dapat dikembangkan dengan strategi yang lebih terkontrol dan adaptif. Informasi topik sebaiknya digunakan sebagai fitur tambahan atau pembobot kontekstual, bukan sebagai dasar promosi label secara langsung. Pendekatan ini memungkinkan model tetap

memperoleh manfaat pengayaan semantik untuk mendeteksi entitas minor, namun dengan risiko aktivasi berlebihan yang lebih terkendali. Selain itu, eksplorasi metode *topic modeling* yang lebih kontekstual atau *neural topic modeling* dapat menjadi alternatif untuk meningkatkan kualitas representasi tematik.

Ketiga, pengembangan representasi struktural melalui GCN dapat diperluas dengan eksplorasi variasi arsitektur graf, seperti *Graph Attention Network* (GAT) atau mekanisme pembobotan relasi yang lebih adaptif. Mengingat hasil penelitian menunjukkan bahwa GCN memberikan keseimbangan terbaik antara sensitivitas dan kontrol kesalahan, pengembangan pada sisi pemodelan graf berpotensi meningkatkan performa tanpa memperbesar propagasi kesalahan secara sistematis.

Keempat, penelitian selanjutnya dapat mengeksplorasi strategi penanganan ketidakseimbangan label yang lebih beragam, seperti *focal loss* yang lebih adaptif, *dynamic class weighting*, atau teknik *data augmentation* untuk entitas minor. Pendekatan ini dapat membantu meningkatkan kemampuan model dalam mengenali entitas yang jarang muncul tanpa meningkatkan prediksi keliru pada kelas non-entitas.

Kelima, evaluasi lanjutan dapat dilakukan pada variasi jenis teks yang lebih beragam, seperti isi berita lengkap atau sumber berita lintas platform. Pengujian pada domain yang lebih luas akan memberikan gambaran mengenai kemampuan generalisasi model di luar karakteristik judul berita yang relatif pendek dan padat.

Terakhir, dalam implementasi praktis, pemilihan model NER perlu disesuaikan dengan kebutuhan sistem. Model IndoBERT–BiLSTM sesuai digunakan pada aplikasi yang memprioritaskan kestabilan dan ketepatan prediksi. Model berbasis LDA relevan pada sistem yang menekankan sensitivitas tinggi terhadap isu kesehatan dan meminimalkan kehilangan entitas penting. Sementara itu, model dengan integrasi GCN dapat dipertimbangkan sebagai pendekatan yang lebih seimbang untuk sistem yang membutuhkan peningkatan deteksi entitas minor dengan tetap menjaga konsistensi dan stabilitas prediksi secara keseluruhan.

DAFTAR PUSTAKA

Amazinum. 2023. What is NLP and how it is implemented in our lives? Amazinum Blog. https://amazinum.com/insights/what-is-nlp-and-how-it-is-implemented-in-our-lives/.

Amien, M. dan Gunawan, G. F. 2024. BERT dan bahasa Indonesia: studi tentang efektivitas model NLP berbasis transformer. *ELANG: Journal of Interdisciplinary Research*. 1(2): 132–140.

Amien, M. 2023. Sejarah dan perkembangan teknik natural language processing (NLP) bahasa Indonesia: tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia. *arXiv preprint*. arXiv:2304.02746. https://doi.org/10.48550/arXiv.2304.02746.

Arai, K. Oda, M. dan Sato, T. 2023. Method for hyperparameter tuning of EfficientNetV2-based image classification by deliberately modifying Optuna tuned result. *International Journal of Advanced Computer Science & Applications*. 14(12): 1–10. https://doi.org/10.14569/IJACSA.2023.0141248.

Archana, S. M. dan Prakash, J. 2024. Biomedical named entity recognition through improved balanced undersampling for addressing class imbalance and preserving contextual information. *International Journal of Information Technology*. 16(8): 4995–5003. https://doi.org/10.1007/s41870-024-02137-w.

Baker, M. A. Sands, K. E. Huang, S. S. Kleinman, K. Septimus, E. J. Varma, N. Blanchard, J. Poland, R. E. Coady, M. H. Yokoe, D. S. dan lainnya. 2022. The impact of coronavirus disease 2019 (COVID-19) on healthcare-associated infections. *Clinical Infectious Diseases*. 74(10): 1748–1754. https://doi.org/10.1093/cid/ciab688.

Batool, A. dan Byun, Y.-C. 2024. Enhanced sentiment analysis and topic modeling during the pandemic using automated latent Dirichlet allocation. *IEEE Access*. 12: 81206–81220. https://doi.org/10.1109/ACCESS.2024.3411717.

Belbekri, A. Bouarroudj, W. dan Benchikha, F. 2024. A two-stage GAN oversampling: integrating GPT-3 and DBpedia for named entity recognition datasets. *Proceedings of the TACC 2024 Conference*. https://ceur-ws.org/Vol-3973/paper6.pdf.

Benchama, A. dan Zebbara, K. 2024. Fine-tuning CNN-BiGRU for intrusion detection with SMOTE optimization using Optuna. *Salud, Ciencia y Tecnología - Serie de Conferencias*. (3): 968. https://doi.org/10.56294/sctconf2024968.

Bin Naeem, S. dan Boulos, M. N. K. 2021. COVID-19 misinformation online and health literacy: a brief overview. *International Journal of Environmental Research and Public Health*. 18(15): 8091. https://doi.org/10.3390/ijerph18158091.

Blei, D. M., Ng, A. Y., dan Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 3(Jan): 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Budi, I. dan Suryono, R. R. 2023. Application of named entity recognition method for Indonesian datasets: a review. *Bulletin of Electrical Engineering and Informatics*. 12(2): 969–978. https://doi.org/10.11591/eei.v12i2.4529.

Buenano-Fernandez, D. González, M. Gil, D. dan Luján-Mora, S. 2020. Text mining of open-ended questions in self-assessment of university teachers: an LDA topic modeling approach. *IEEE Access*. 8: 35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983.

Castaño-Pulgarín, S. A. Suárez-Betancur, N. Vega, L. M. T. dan López, H. M. H. 2021. Internet, social media and online hate speech: systematic review. *Aggression and Violent Behavior*. 58: 101608. https://doi.org/10.1016/j.avb.2021.101608.

Chen, Q. dan Shen, Y. 2025. NER based on dependency structure feature fusion. *Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network*. 684–691. https://doi.org/10.1145/3727353.3727463.

Courant, R. Edberg, M. Dufour, N. dan Kalogeiton, V. 2023. Transformers and visual transformers. *Machine Learning for Brain Disorders*. 193–229. https://arxiv.org/pdf/2303.12068.

De Magistris, G. Russo, S. Roma, P. Starczewski, J. T. dan Napoli, C. 2022. An explainable fake news detector based on named entity recognition and stance classification applied to COVID-19. *Information*. 13(3): 137. https://doi.org/10.3390/info13030137.

Deng, N. Fu, H. dan Chen, X. 2021. Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF. *Wireless Communications and Mobile Computing*. 2021: 6696205. https://doi.org/10.1155/2021/6696205.

Devlin, J. Chang, M.-W. Lee, K. dan Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805.

Excler, J.-L. Saville, M. Berkley, S. dan Kim, J. H. 2021. Vaccine development for emerging infectious diseases. *Nature Medicine*. 27(4): 591–600. https://doi.org/10.1038/s41591-021-01301-0.

Evtimova, M. 2023. Hyperparameter tuning for address validation using Optuna. *WSEAS Transactions on Computer Research*. 12: 105–111. https://doi.org/10.37394/232018.2024.12.10.

Gao, W. Zheng, X. dan Zhao, S. 2021. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF. *Journal of Physics: Conference Series*. 1848(1): 012083. https://doi.org/10.1088/1742-6596/1848/1/012083.

Gangadharan, V. dan Gupta, D. 2020. Recognizing named entities in agriculture documents using LDA based topic modelling techniques. *Procedia Computer Science*. 171: 1337–1345. https://doi.org/10.1016/j.procs.2020.04.143.

Gardazi, N. M. Daud, A. Malik, M. K. Bukhari, A. Alsahfi, T. dan Alshemaimri, B. 2025. BERT applications in natural language processing: a review. *Artificial Intelligence Review*. 58(6): 1–49. https://doi.org/10.1007/s10462-025-11162-5.

Hanh, T. T. H. Doucet, A. Sidere, N. Moreno, J. G. dan Pollak, S. 2021. Named entity recognition architecture combining contextual and global features. *International Conference on Asian Digital Libraries*. 264–276. https://doi.org/10.1007/978-3-030-91669-$5_2$1.

Haque, M. Z. Zaman, S. Saurav, J. R. Haque, S. Islam, M. S. dan Amin, M. R. 2023. B-NER: a novel Bangla named entity recognition dataset with largest entities and its baseline evaluation. *IEEE Access*. 11: 45194–45205. https://doi.org/10.1109/ACCESS.2023.3267746.

Heydarian, M. Doyle, T. E. dan Samavi, R. 2022. MLCM: multi-label confusion matrix. *IEEE Access*. 10: 19083–19095. https://doi.org/10.1109/ACCESS.2022.3151048.

Hicks, S. A. Strümke, I. Thambawita, V. Hammou, M. Riegler, M. A. Halvorsen, P. dan Parasa, S. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*. 12(1): 5979. https://doi.org/10.1038/s41598-022-09954-8.

Hochreiter, S. dan Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*. 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Ilemobayo, J. A. Durodola, O. Alade, O. Awotunde, O. J. Olanrewaju, A. T. Falana, O. Ogungbire, A. Osinuga, A. Ogunbiyi, D. Ifeanyi, A. dan lainnya. 2024. Hyperparameter tuning in machine learning: a comprehensive review. *Journal of Engineering Research and Reports*. 26(6): 388–395. https://doi.org/10.9734/jerr/2024/v26i61188.

Ivanenko, V. 2025. Hybrid model for financial named entity recognition in Ukrainian using CRF, BiLSTM, and BERT. *WSEAS Transactions on Systems*. 24: 569–581. https://doi.org/10.37394/23202.2025.24.50.

Jagwani, V. Meghani, S. Pai, K. dan Dhage, S. 2023. Resume evaluation through latent Dirichlet allocation and natural language processing for effective candidate selection. *arXiv preprint*. arXiv:2307.15752. https://doi.org/10.48550/arXiv.2307.15752.

Jehangir, B. Radhakrishnan, S. dan Agarwal, R. 2023. A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal*. 3: 100017. https://doi.org/10.1016/j.nlp.2023.100017.

Ju, W. Sannusi, S. N. dan Mohamad, E. 2023. Stigmatizing monkeypox and COVID-19: a comparative framing study of The Washington Post's online news. *International Journal of Environmental Research and Public Health*. 20(4): 3347. https://doi.org/10.3390/ijerph20043347.

Kee, T. dan Ho, W. K. O. 2025. Optimizing machine learning models for urban sciences: a comparative analysis of hyperparameter tuning methods. *Urban Science*. 9(9): 348. https://doi.org/10.3390/urbansci9090348.

Khairunnisa, S. O. Chen, Z. dan Komachi, M. 2023. Dataset enhancement and multilingual transfer for named entity recognition in the Indonesian language. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 22(6): 1–21. https://doi.org/10.1145/3592854.

Khanfir, Y. Dhiaf, M. Ghodhbani, E. Rouhou, A. C. dan Kessentini, Y. 2024. Graph neural networks for end-to-end information extraction from handwritten documents. *WACV 2024*. 504–512. https://doi.org/10.1109/WACV57701.2024.00056.

Kipf, T. N. dan Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint*. arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907.

Kochnev, R. Goodarzi, A. T. Bentyn, Z. A. Ignatov, D. dan Timofte, R. 2025. Optuna vs Code Llama: are LLMs a new paradigm for hyperparameter tuning? *arXiv preprint*. arXiv:2504.06006. https://doi.org/10.48550/arXiv.2504.06006.

Krawczyk, K. Chelkowski, T. Laydon, D. J. Mishra, S. Xifara, D. Gibert, B. Flaxman, S. Mellan, T. Schwämmle, V. Röttger, R. dan lainnya. 2021. Quantifying online news media coverage of the COVID-19 pandemic: text mining study and resource. *Journal of Medical Internet Research*. 23(6): e28253. https://doi.org/10.2196/preprints.31544.

Lee, L.-H. Lu, C.-H. dan Lin, T.-M. 2022. NCUEE-NLP at SemEval-2022 Task 11: Chinese named entity recognition using the BERT-BiLSTM-CRF model. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 1597–1602. https://doi.org/10.18653/v1/2022.semeval-1.220.

Li, Q. L. Yan, S. J. Chen, Q. dan Zhang, K. 2024. Research on Chinese named entity recognition based on BERT-CNN-BiLSTM-CRF model with fusion multi-head attention mechanism. *2024 14th International Conference on Information Science and Technology (ICIST)*. 583–588. https://doi.org/10.1109/ICIST63249.2024.10805339.

Lopez, P. Du, C. Cohoon, J. Ram, K. dan Howison, J. 2021. Mining software entities in scientific literature: document-level NER for an extremely imbalanced and large-scale task. *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*. 3986–3995. https://doi.org/10.1145/3459637.3481936.

Mach, K. J. Salas Reyes, R. Pentz, B. Taylor, J. Costa, C. A. Cruz, S. G. Thomas, K. E. Arnott, J. C. Donald, R. Jagannathan, K. dan lainnya. 2021. News media coverage of COVID-19 public health and policy information. *Humanities and Social Sciences Communications*. 8(1): 30. https://doi.org/10.1057/s41599-021-00900-z.

Naik, D. dan Jaidhar, C. D. 2022. A novel multi-layer attention framework for visual description prediction using bidirectional LSTM. *Journal of Big Data*. 9(1): 104. https://doi.org/10.1186/s40537-022-00664-6.

Nabiilah, G. Z. Alam, I. N. Purwanto, E. S. dan Hidayat, M. F. 2024. Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical & Computer Engineering*. 14(1): 1071–1078. https://doi.org/10.11591/ijece.v14i1.pp1071-1078.

Nemoto, S. Kitada, S. dan Iyatomi, H. 2024. Majority or minority: data imbalance learning method for named entity recognition. *IEEE Access*. https://doi.org/10.48550/arXiv.2401.11431.

Noh, J. dan Kavuluru, R. 2021. Joint learning for biomedical NER and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 1–10. https://doi.org/10.1145/3459930.3469533.

Nuryana, I. K. D. Mawarni, L. I. D. dan Juanara, E. 2025. Early detection of environmental issues from social media using IndoBERT and LDA: case study of pollution and deforestation in Indonesia. *E3S Web of Conferences*. 645: 05005. https://doi.org/10.1051/e3sconf/202564505005.

Nuryanto, N. 2025. Deteksi berita hoaks berbahasa Indonesia menggunakan natural language processing (NLP) dengan model IndoBERT dan implementasi berbasis web. *Prosiding Seminar Nasional Universitas Ma Chung (Informatika & Sistem Informasi; Bahasa dan Seni; Farmasi)*. 5(1).

Purnomo, T. D. dan Sutopo, J. 2024. Comparison of pre-trained BERT-based transformer models for regional language text sentiment analysis in Indonesia. *International Journal of Science and Technology*. 3(3): 11–21. https://doi.org/10.56127/ijst.v3i3.1739.

Rakhmawati, N. A. Cisatra, A. Ansori, D. D. M. Akmal, D. N. F. A. dan Ramadhani, S. 2024. Identifikasi topik hangat di media berita menggunakan latent

Dirichlet allocation. *JIEET (Journal of Information Engineering and Educational Technology)*. 8(1): 14–17. https://doi.org/10.26740/jieet.v8n1.p14-17.

Rocha, Y. M. De Moura, G. A. Desidério, G. A. De Oliveira, C. H. Lourenço, F. D. dan de Figueiredo Nicolete, L. D. 2023. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health*. 31(7): 1007–1016. https://doi.org/10.1007/s10389-021-01658-z.

Santoso, J. Setiawan, E. I. Purwanto, C. N. Yuniarno, E. M. Hariadi, M. dan Purnomo, M. H. 2021. Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory. *Expert Systems with Applications*. 176: 114856. https://doi.org/10.1016/j.eswa.2021.114856.

Sarantopoulos, A. Mastori Kourmpani, C. Yokarasa, A. L. Makamanzi, C. Antoniou, P. Spernovasilis, N. dan Tsioutis, C. 2024. Artificial intelligence in infectious disease clinical practice: an overview of gaps, opportunities, and limitations. *Tropical Medicine and Infectious Disease*. 9(10): 228. https://doi.org/10.3390/tropicalmed9100228.

Sathyanarayanan, S. dan Tantri, B. R. 2024. Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*. 27(4S): 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345.

Sayarizki, P. dan Nurrahmi, H. 2024. Implementation of IndoBERT for sentiment analysis of Indonesian presidential candidates. *Indonesian Journal on Computing*. 9(2): 61–72. https://doi.org/10.34818/INDOJC.2024.9.2.934.

Setiawan, E. I. Dharmawan, W. Halim, K. J. Santoso, J. Ferdinandus, F. X. Fujisawa, K. dan Purnomo, M. H. 2024. Indonesian news stance classification based on hybrid bidirectional LSTM and transformer-based embedding. *International Journal of Intelligent Engineering & Systems*. 17(5): 517–537. https://doi.org/10.22266/ijies2024.1031.41.

Shah, S. A. A. Masood, M. A. dan Yasin, A. 2022. Dark web: e-commerce information extraction based on named entity recognition using bidirectional-LSTM. *IEEE Access*. 10: 99633–99645. https://doi.org/10.1109/ACCESS.2022.3206539.

Sharma, S. Datta, A. Shankaran, V. dan Sharma, R. 2023. Misinformation concierge: a proof-of-concept with curated Twitter dataset on COVID-19 vaccination. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 5091–5095. https://doi.org/10.48550/arXiv.2309.00639.

Shidik, G. F. Saputra, F. O. Saraswati, G. W. Winarsih, N. A. S. Rohman, M. S. Pramunendar, R. A. Kusuma, E. J. Ratmana, D. O. Venus, V. dan Andono, P. N. 2024. Indonesian disaster named entity recognition from multi-source information using bidirectional LSTM (BiLSTM). *Journal of Open Innovation: Technology, Market, and Complexity*. 10(3): 100358. https://doi.org/10.1016/j.joitmc.2024.100358.

Subowo, E., Bukhori, I., dan Warto. 2025. Corpus development and NER model for identification of legal entities (articles, laws, and sanctions) in corruption court decisions in Indonesia. *Transactions on Informatics and Data Science*, 2(1), 27–39. https://doi.org/10.24090/tids.v2i1.13592.

Taher, H. A. Hasan, N. N. A. dan Mahdi, B. 2025. Integration named entity recognition and latent Dirichlet allocation to enhance topic modeling. *Annals of Emerging Technologies in Computing (AETiC)*. 9(2). https://doi.org/10.33166/AETiC.2025.02.002.

Umam, A. K. Alzami, F. Sani, R. R. Rohmani, A. Prabowo, D. P. Pergiwati, D. Megantara, R. A. dan Iswahyudi, I. 2025a. Enhancing entity extraction in e-government complaint data using LDA-assisted NER. *Sinkron: Jurnal dan Penelitian Teknik Informatika*. 9(4): 1878–1888. https://doi.org/10.33395/sinkron.v9i4.15292.

Umam, M. Z. Putra, H. A. dan Wibowo, A. 2025b. Semantic pre-annotation for named entity recognition using topic modeling. *Journal of e-Government Studies and Research*. 8(1): 44–55.

Vaswani, A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A. N. Kaiser, Ł. dan Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*. 30. https://doi.org/10.48550/arXiv.1706.03762.

Vujović, Ž. 2021. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*. 12(6): 599–606. https://doi.org/10.14569/IJACSA.2021.0120670.

Wang, T. Zhang, Y. Zhang, Y. Lu, H. Yu, B. Peng, S. Ma, Y. dan Li, D. 2023a. A hybrid model based on deep convolutional network for medical named entity recognition. *Journal of Electrical and Computer Engineering*. 2023(1): 8969144. https://doi.org/10.1155/2023/8969144.

Wang, S. Meng, Y. Ouyang, R. Li, J. Zhang, T. Lyu, L. dan Wang, G. 2023b. GNN-SL: sequence labeling based on nearest examples via GNN. *Findings of the Association for Computational Linguistics: ACL 2023*. 12679–12692. https://doi.org/10.18653/v1/2023.findings-acl.803.

Wang, J. Huang, J. X. Tu, X. Wang, J. Huang, A. J. Laskar, M. T. R. dan Bhuiyan, A. 2024. Utilizing BERT for information retrieval: survey, applications, resources, and challenges. *ACM Computing Surveys*. 56(7): 1–33. https://doi.org/10.1145/3648471.

Wang, T. Xu, Y. Qin, Y. Wang, X. Zheng, F. dan Li, W. 2025a. Short-term PV forecasting of multiple scenarios based on multi-dimensional clustering and hybrid transformer-BiLSTM with ECPO. *Energy*. 137654. https://doi.org/10.1016/j.energy.2025.137654.

Wang, C. Dong, Q. Wang, X. dan Sui, Z. 2025b. Statistical dataset evaluation: a case study on named entity recognition. *Natural Language Processing*. 31(1): 90–110. https://doi.org/10.1017/nlp.2024.37.

Xu, L. dan Li, J. 2021. Biomedical named entity recognition based on BERT and BiLSTM-CRF. *Computer Engineering and Science*. 43(10): 1873–1880. http://joces.nudt.edu.cn/EN/Y2021/V43/I10/1873.

Yulianti, E. Bhary, N. Abdurrohman, J. Dwitilas, F. W. Nuranti, E. Q. dan Husin, H. S. 2024. Named entity recognition on Indonesian legal documents: a dataset and study using transformer-based models. *International Journal of Electrical and Computer Engineering*. 14(5): 5489–5501. https://doi.org/10.11591/ijece.v14i5.pp5489-5501.

Zainuddin, Z. dan Tahir, Z. 2025. Entity extraction in Indonesian online news using named entity recognition (NER) with hybrid method transformer, Word2Vec, attention and Bi-LSTM. *JOIV: International Journal on Informatics Visualization*. 9(3): 964–973. https://doi.org/10.62527/joiv.9.3.2902.

Zaratiana, U. Tomeh, N. Holat, P. dan Charnois, T. 2022. GNNer: reducing overlapping in span-based NER using graph neural networks.

Proceedings of the ACL Student Research Workshop. 97–103. https://doi.org/10.18653/v1/2022.acl-srw.9.

Zhang, J. 2021. Combining GCN and transformer for Chinese grammatical error detection. *arXiv preprint*. arXiv:2105.09085. https://doi.org/10.53106/160792642022122307020.

Zhou, L. Wang, T. Qu, H. Huang, L. dan Liu, Y. 2020. A weighted GCN with logical adjacency matrix for relation extraction. *ECAI 2020*. 2314–2321. https://doi.org/10.3233/FAIA200360.

Zhou, Y. Zeng, H. Zhang, W. dan Liu, J. 2024. Named entity recognition model based on multi-BiLSTM and competition mechanism. *Electronics Letters*. 60(9): e13194. https://doi.org/10.1049/ell2.13194.