TESIS

EKSTRAKSI ENTITAS PADA BERITA ONLINE BAHASA INDONESIA MENGGUNAKAN *NAMED ENTITY RECOGNITION (NER)* DENGAN METODE *HYBRID BI-LSTM* DAN *TRANSFORMER*

Entity Extraction in Indonesian Online News Using Named Entity Recognition (NER) with Hybrid Bi-LSTM and Transformer Methods

MUDASSIR
D082211007

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA
DEPERTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

PENGAJUAN TESIS

EKSTRAKSI ENTITAS PADA BERITA ONLINE BAHASA INDONESIA MENGGUNAKAN *NAMED ENTITY RECOGNITION (NER)* DENGAN METODE *HYBRID BI-LSTM* DAN *TRANSFORMER*

Tesis
Sebagai Salah Satu Syarat Untuk Mencapai Gelar Magister
Program Studi Teknik Informatika

Disusun dan diajukan oleh

MUDASSIR
D082211007

Kepada

FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

TESIS

EKSTRAKSI ENTITAS PADA BERITA ONLINE BAHASA INDONESIA MENGGUNAKAN *NAMED ENTITY RECOGNITION (NER)* DENGAN METODE *HYBRID BI-LSTM* DAN *TRANSFORMER*

MUDASSIR
D082211007

Telah dipertahankan di hadapan Panitia Ujian Tesis yang dibentuk dalam rangka penyelesaian studi pada
Program Magister Teknik Informatika Fakultas Teknik
Universitas Hasanuddin
Pada tanggal 20 November 2024
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama

Dr. Ir. Zahir Zainuddin, M. Sc
NIP. 19640427 198910 1 002

Pembimbing Pendamping

Dr.Eng. Zulkifli Tahir, S.T., M.Sc.
NIP. 19840403 201012 1 004

Dekan Fakultas Teknik
Universitas Hasanuddin

Prof. Dr.Eng. Ir. Muhammad Isran Ramli, M.T. IPM., ASEAN.Eng.
NIP. 19730926 200012 1 002

Ketua Program Studi
S2 Teknik Informatika

Dr. Ir. Zahir Zainuddin, M.Sc.
NIP. 19640427 198910 1 002

PERNYATAAN KEASLIAN TESIS DAN PELIMPAHAN HAK CIPTA

Yang bertanda tangan di bawah ini

Nama	: Mudassir
Nomor mahasiswa	: D082211007
Program studi	: Magister Teknik Informatika

Dengan ini menyatakan bahwa, tesis yang berjudul "EKSTRAKSI ENTITAS PADA BERITA ONLINE BAHASA INDONESIA MENGGUNAKAN *NAMED ENTITY RECOGNITION (NER)* DENGAN METODE *HYBRID BI-LSTM* DAN *TRANSFORMER*" adalah benar karya saya dengan arahan dari komisi pembimbing Dr. Ir. Zahir Zainuddin, M.Sc. dan Dr.Eng. Zulkifli Tahir, S.T., M.Sc. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apapun kepada perguruan tinggi manapun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini.

Dengan ini saya melimpahkan hak cipta ini dari karya tulis saya berupa tesis ini kepada Universitas Hasanuddin.

Gowa, November 2024

Yang menyatakan

Mudassir

KATA PENGANTAR

Alhamdulillahi rabbil'alamin, segala puji bagi Allah Subhanahu Wa Ta'ala Yang Maha Sempurna, yang telah memberikan rahmat, hidayah dan pertolongan-Nya sehingga penulis dapat menyelesaikan tesis dengan judul "Ekstraksi entitas pada teks berita online Bahasa Indonesia menggunakan Named Entity Recognition (NER) dengan metode *hybrid Bi-LSTM* dan *Transformer*". Tak lupa pula shalawat dan salam kepada Nabi Muhammad Shallahu 'Alaihi Wasallam yang telah menyinari dunia ini dengan keindahan ilmu dan akhlak yang diajarkan kepada seluruh umatnya.

Tesis ini disusun untuk memenuhi persyaratan untuk memperoleh gelar Magister Komputer (M.Kom.) pada Program Pascasarjana Departemen Teknik Informatika Universitas Hasanuddin Makassar. Tentunya penyelesaian tesis ini tidak terlepas dari dukungan dan bantuan dari semua pihak. Untuk itu, dengan penuh kerendahan hati penulis menyampaikan terima kasih setulus-tulusnya dan setinggi-tingginya kepada:

1. Ayahanda penulis Massi, S.Pd., M.Si. dan ibunda tercinta Yajirah, S.Pd. dan adik saya satu-satunya Nurmayanti yang telah memberikan dukungan materil, doa dan motivasi yang kuat kepada penulis, hingga penulis dapat menyelesaikan penelitian ini.
2. Bapak Dr. Ir. Zahir Zainuddin, M.Sc. sebagai pembimbing pertama sekaligus sebagai Ketua Program Studi S2 Teknik Informatika dan Bapak Dr.Eng. Zulkifli Tahir, S.T., M.Sc. selaku dosen pembimbing kedua yang telah meluangkan waktunya kepada penulis untuk membimbing, memberikan masukan, memotivasi tiada henti-hentinya hingga tahap penyelesaian tesis ini.
3. Rekan-rekan Lab *Computer Based System* Teknik Informatika yang selalu saling mendukung dalam suka maupun duka dalam proses penyelesaian tesis ini.
4. Rekan–rekan Mahasiswa S2 Teknik Informatika Angkatan 2021 yang selalu mendukung dalam proses penyelesaian tesis ini.

Penulis menyadari bahwa tesis masih jauh dari kata sempurna dan di dalam penyelesaiannya masih menemui kesulitan dan hambatan, sehingga penulis tetap mengharapkan saran dan kritik untuk pengembangan lebih lanjut, agar dapat memberikan manfaat yang banyak bagi semua pembaca.

Gowa, 5 November 2024

Mudassir

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Mudassir. *Ekstraksi entitas pada teks berita online bahasa Indonesia menggunakan Named Entity Recognition (NER) dengan metode hybrid Bi-LSTM dan Transformer* (dibimbing oleh **Zahir Zainuddin** dan **Zulkifli Tahir**)

Named Entity Recognition (NER) adalah domain penting dalam *Natural Language Processing (NLP)* yang mengidentifikasi entitas seperti nama orang, lokasi, dan organisasi dalam teks. Meskipun banyak penelitian *NER* yang berfokus pada bahasa Inggris, masih diperlukan lebih banyak penelitian tentang *NER* bahasa Indonesia. Indonesia memiliki tantangan yang unik karena kompleksitas dan ambiguitas strukturalnya. Pembelajaran mesin konvensional dan teknik pembelajaran mendalam telah digunakan dalam NER, tetapi mengintegrasikan metode-metode ini untuk meningkatkan kinerja masih belum dieksplorasi secara rinci. Penelitian ini menyajikan model hibrida baru, yang menggabungkan mekanisme Transformer dan Bidirectional Long Short-Term Memory (Bi-LSTM untuk meningkatkan performa *NER* pada teks bahasa Indonesia. Transformer dan Bi-LSTM memanfaatkan keunggulan masing-masing komponen untuk menghasilkan representasi vektor kata yang lebih baik, mengekstrak fitur kalimat yang rumit, dan mengacaukan entitas secara kontekstual. Eksperimen kami menunjukkan keefektifan model hibrida yang diusulkan, yang mencapai peningkatan signifikan dalam kinerja *NER*. Secara khusus, metide yang diusulkan mencapai F1-Score 81,00 pada dataset berita online Indonesia, melampaui model *Bi-LSTM* tradisional, yang mencapai skor 76,11. Hasil penelitian menunjukkan bahwa *Transformer* dan *Bi-LSTM* secara efektif mengurangi ambiguitas dan menangkap konteks, sehingga meningkatkan akurasi pengenalan entitas. Penelitian di masa depan harus berfokus pada pengurangan waktu komputasi untuk dataset yang lebih besar tanpa mengorbankan kinerja *NER* secara keseluruhan. Penelitian ini menggarisbawahi potensi mengintegrasikan teknik pembelajaran mendalam yang canggih untuk mengatasi tantangan unik *NER* Indonesia, memberikan landasan untuk kemajuan lebih lanjut di bidang ini.

Kata Kunci: Named Entity Recognition, *Transformer*, Bahasa Indonesia

ABSTRACT

Mudassir. *Entity Extraction in Indonesian Online News Using Named Entity Recognition (NER) With Hybrid Bi-LSTM and Transformer Methods* **(**supervised by **Zahir Zainuddin** dan **Zulkifli Tahir)**

Named Entity Recognition (NER) is a vital field in Natural Language Processing (NLP) that recognises entities such as names of people, places, and organisations in text. A lot of NER research uses models applied to the English language, but there are still few in Indonesian. At the same time, Indonesian is also a language prone to ambiguity in determining its entities due to its complexity and several structural differences from English. So far, NER development rarely uses deep learning and combines it, even though combining deep learning models such as *Transformer*, Word2Vec, Attention, and *Bi-LSTM* (TWBiL) shows the potential to improve NER performance in Indonesian. TWBiL combines *Transformer*, Word2Vec, *Bi-LSTM* and attention mechanisms to obtain a better word vector representation, extract sentence features, and pay attention to context to reduce possible ambiguities during detection. In this study, the proposed combination of methods succeeded in performing named entity extraction well, and the results were significant. The evaluation outcomes reveal that TWBiL attained an F1-Score of 81,00 on the Indonesian online news dataset, surpassing the *Bi-LSTM* (Bidirectional Long Short-Term Memory) model, which scored 76,18. Future research should focus on reducing computation time for larger datasets without compromising the overall performance of NER.

Keywords: Named Entity Recognition, *Natural Language Processing*, Bahasa Indonesia

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pengenalan entitas bernama adalah salah satu topik penelitian dalam NLP (*Natural Language Processing*) yang bertugas mengidentifikasi dan mengklasifikasikan entitas bernama dalam teks. Fokus utama dalam pengembangan NER adalah menemukan model yang lebih akurat dan tangguh untuk berbagai domain dan bahasa. NER pertama kali digunakan ke dalam Message Understanding Conference-6 (MUC-6) pada tahun 1995. Penerapan NER juga banyak dilakukan pada aplikasi pemrosesan bahasa alami seperti sistem tanya jawab otomatis, pengindeksan dokumen, mesin penerjemah, pencarian informasi, dan peringkasan teks [1]. Tugas dari NER adalah untuk mengidentifikasi atau mengklasifikasi sebuah entitas misalnya nama orang, organisasi, lokasi dan sesuatu entitas lain dalam sebuah teks yang sangat berguna dalam kasus ekstraksi informasi. Sebagai contoh, pada kalimat berikut: “CEO Apple Tim Cook melakukan demo terkait rilis produk terbarunya di Sillicon Valley. Demo tersebut diselenggarakan di Apple Campus, California”, NER akan mengenali “Tim Cook” sebagai nama orang, “Apple” sebagai nama organisasi/perusahaan, dan “Apple Campus” sebagai nama lokasi. Salah satu kekurangan utama pada NER termasuk pada bahasa Indonesia adalah ambiguitas. Pada contoh di atas misalnya, kemunculan kata “Apple” pertama memiliki arti berbeda dengan kemunculan kata “Apple” kedua. Kata pertama adalah nama sebuah organisasi/perusahaan sedangkan kata kedua adalah nama sebuah gedung. Manusia memiliki kemampuan yang mampu membedakan arti kedua kata tersebut, tetapi tidak demikian dengan program komputer. Oleh karena itu dibutuhkan NER agar komputer dapat melakukan pengidentifikasian terhadap entitas.

Sebagian besar penelitian NER saat ini menggunakan bahasa Inggris karena ketersediaan dataset anotasi berkualitas tinggi. Banyak korpus bahasa Inggris yang luas dan tersedia untuk publik, seperti CoNLL-2003, JNLPBA, dan GENIA, NCBI, yang memungkinkan peneliti melatih dan mengevaluasi model NER dengan akurasi

tinggi. Bahasa Inggris juga dominan di berbagai bidang seperti bisnis, sains, dan teknologi, sehingga permintaan untuk sistem NER bahasa Inggris lebih tinggi. Akibatnya, ada pasar dan insentif lebih besar untuk pengembangan dan peningkatan model NER bahasa Inggris.

Bahasa Indonesia adalah salah satu bahasa dengan sedikit sumber daya dalam topik penelitian NER. Penelitian awal NER dalam Bahasa Indonesia dilakukan oleh [2] menggunakan pendekatan berbasis aturan dengan metode maksimum entropi pada korpus berita dari surat kabar harian nasional online. Tahun-tahun berikutnya, suda ada beberapa peneliti dan akademisi mengembangkan NER dalam Bahasa Indonesia, termasuk [3, 4, 5, 6]. Aryoyudanta menggunakan pendekatan semi-terawasi dengan algoritma co-training [6], sementara R. A. Leonandya mengusulkan algoritma semi-terawasi untuk NER [7]. Gunawan [8] menggunakan model LSTM-CNN, Azalia [9] menggunakan klasifikasi Naïve Bayes untuk pengindeksan nama dalam terjemahan hadits Indonesia, dan Fu mengusulkan metode fitur berbasis perhatian terstruktur hierarkis untuk NER [10]. Berbagai korpus dikembangkan dari banyak sumber seperti artikel berita [6], Wikipedia [11][7], dan teks agama [4]. Beberapa penelitian menggunakan dataset khusus seperti terjemahan hadits di Azalia [4], yang membatasi penerapannya ke domain lain.

Rachman [3] berfokus pada NER pada postingan Twitter Indonesia menggunakan jaringan LSTM. Mereka menggunakan dataset yang terdiri dari 480 teks berita yang diberi label dengan tiga jenis entitas: Orang, Lokasi, dan Organisasi. Eksperimen menunjukkan bahwa jaringan LSTM mengungguli model baseline lainnya dengan skor F1 sebesar 77,08%. Setiyoaji [12] mengusulkan pendekatan berbasis model Markov tersembunyi (HMM) dan algoritma Viterbi untuk NER pada penerjemahan Al-Quran ke dalam bahasa Indonesia dengan skor F1 sebesar 72,55%.

D'Souza membahas tantangan NER dalam Memeriksa sumber daya bahasa NER CS yang ada dan menjelaskan masalah kurangnya entitas standar di dalamnya [13]. Kombinasi BiLSTM + CRF menunjukkan skor f1 sebesar 72,62, sedangkan BiLSTM + CNN + CRF mencapai 75,18. Menggabungkan sumber daya yang ada dengan memberikan anotasi data tambahan dapat menciptakan korpus besar yang

tersedia untuk umum. Lai [14] mengusulkan pendekatan berbasis *Transformer* yang kompetitif dan menempati peringkat ke-12 dari 30 tim dengan skor F1 makro sebesar 72,50%. Pendekatan augmentasi data menggunakan pengaitan entitas tidak meningkatkan kinerja sistem secara keseluruhan.

Beberapa pendakatan yang dapat dilakukan untuk mengenali entitas yaitu menggunakan rule-base, deep learning, dan deep learning. Salah satu metode NLP yang bisa dilakukukan adalah dengan menggunakan deep learning dengan metode Named Entity Rocognition (NER). Pada penelitian-penelitian terbaru mengenai NER telah dilakukan juga penelitian dengan menggunakan metode *hybrid*. Pendekatan *Hybrid* telah diimplementasikan dalam Tugas Pemrosesan Bahasa Alami. Model hibrida telah terbukti dapat meningkatkan kinerja berbagai model. Suncong et. al.[15] menggunakan hibrida LSTM dan CNN untuk mendapatkan entitas dan relasinya. Model hibrida lain dalam analisis sentimen adalah menggabungkan Gradient Boosting Decision Tree dan Support Vector Machine dalam [16]. Metode *Hybrid* juga dikembangkan untuk Summarization of microblog posts [17] dan Sentiment Analysis of Political Data [18]. Alasan utama mengapa pendekatan *hybrid* dipilih sebagai metode yang diusulkan adalah karena dapat menggabungkan kekuatan masing-masing model untuk mendapatkan kinerja yang lebih baik.

Penelitian ini mengusulkan sebuah model hibrida untuk mendapatkan Named Entity Indonesia dengan memasukkan Word Embedding ke dalam algoritma dasar. Model *hybrid* ini dibangun dengan menggabungkan word embedding sebagai fitur cluster ke dalam Bidirectional Long ShortTerm Memory (*Bi-LSTM*) seperti penelitian sebelumnya di [19]. Peneliti menggunakan *Transformer* sebagai algoritma clustering. Hasil clustering akan diproses sebagai fitur tambahan selain fitur kata kontekstual standar dan Part-of-Speech (POS) ke dalam *Bi-LSTM* . Kata-kata yang mirip dalam word embedding cenderung memiliki vektor yang mirip. Oleh karena itu, kata-kata yang mirip ini akan dikelompokkan bersama dalam klaster yang sama. Untuk menangkap perilaku ini, peneliti menggunakan fitur cluster dalam *Bi-LSTM* . Hal ini akan membantu *Bi-LSTM* untuk mendapatkan hasil yang lebih baik, seperti hasil dari penelitian sebelumnya di [19]. Jadi, dapat disimpulkan bahwa *Bi-LSTM* dan *Transformer* telah saling melengkapi kekuatan

satu sama lain dalam model hibrida yang diusulkan untuk Pengenalan Entitas Bernama. Penyisipan kata dalam penelitian ini adalah Word2Vec yang telah dilatih sebelumnya berdasarkan penelitian sebelumnya oleh [20]. Beberapa penelitian terkait NER sudah pernah dilakukan dalam berbagai bahasa termasuk Indonesia. Sebuah penelitian[5] yang ditulis oleh Sahrul Sukardi dkk menggunakan metode BiLSTM-CNNs mendapatkan f1 score sebesar 71,37%. Hasil tersebut didapat dari kombinasi metode BiLSTM dan Single CNN. Metode BiLSTM dipilih karena pada penelitian sebelumnya yang menggunakan metode LSTM memiliki kekurangan yaitu pada outputnya yang hanya menerima informasi yang didapat dari masa lampau dan tidak ada akses untuk informasi setelahnya sehingga penentuan entitasnya kurang akurat. Penelitian selanjutnya dengan judul Named-Entity Recognition on Indonesian Teks beritas using Bidirectional LSTM-CRF[3] yang ditulis oleh Deni Cahya Wintakaa dkk. Di dalam penelitian ini penulis membagi pendeteksian entity menjadi dua bagian yaitu pada Bahasa Indonesia formal dan Bahasa Indonesia informal yang datanya diambil dari Twitter. Berdasarkan penelitian ini didapatkan hasil F1 score 86,13% untuk Bahasa Indonesia formal dan 81,17% untuk yang informal.

Pada penelitian lain [13] yang menggunakan metode Bidirectional Long ShortTerm Memory (*Bi-LSTM*) + Part-of-Speech (POS) Tagging yang ditulis oleh Joan Santoso dkk mendapatkan f1 score sebesar 80,18%. Peningkatan akurasi tersebut didapat karena dalam metode *Bi-LSTM* dan Part-of-Speech (POS) Tagging terdapat proses multitask learning yang dapat mengoptimasi algoritma dengan baik dan mengurangi loss function. Penggunaan metode ini juga menutupi kekurangan dari penelitian sebelumnya kerena penggunaan metode Bidirectional LSTM menggabungkan konteks sebelumnya dan konteks setelahnya dengan memproses data dari dua arah. Namun penelitian inipun tidak luput dari kekurangan. Kekurangan dari penelitian ini adalah pada akurasi NER untuk penentuan batas frasa dan hubungan antar kalimat masih kurang sehingga pada proses penentuan entitas oleh NER masih terdapat ambiguitas. Pada penelitian dalam bahasa lain yang ditulis oleh Cillian Berragan dkk meneliti NER untuk penentuan entitas khusus yaitu untuk entitas Geografi dalam hal ini nama tempat. Pada penelitian tersebut memakai metode baru yang masih jarang digunakan pada penelitian NER

dalam bahasa lain termasuk Bahasa Indonesia yaitu metode *Transformer*. Hasil dari performa model NER yang dibuat pada penelitian tersebut menunjukkan adanya peningkatan akurasi pada ekstraksi entitas untuk nama tempat. Namun kekurangan dari metode *Transformer* ini pada proses trainingnya memerlukan data yang cukup besar dan waktu komputasi yang lebih lama.

Berdasarkan penelitian-penelitian sebelumnya didapatkan bahwa pada penelitian NER pada Bahasa Indonesia masih terdapat beberapa kekurangan termasuk masalah ambiguitas pada penentuan entitas sehingga akurasi untuk penentuan entitasnya masih perlu ditingkatkan [13]. Beberapa penelitian yang telah dilakukan trennya menggunakan metode *Bi-LSTM* yang hasilnya memang cukup baik namun berdasarkan penelitian NER dalam bahasa lain terdapat metode lain yang berpotensi untuk meningkatkan akurasi dari NER dalam Bahasa Indonesia dan menutupi kekurangan dari metode *Bi-LSTM* yaitu dengan menggunakan metode *Transformer*. Maka dari itu pada penelitian ini akan digunakan kombinasi dua metode yaitu *Bi-LSTM* dan *Transformer* yang diharapakan dapat memecahkan masalah utama dari NER yaitu ambiguitas, sehingga hasil akhir dari akurasi penentuan entitas dapat ditingkatkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, maka rumusan masalah pada penelitian ini adalah:

1. Bagaimana menganalis ekstraksi entitas pada teks berita online Bahasa Indonesia menggunakan *Named Entity Recognition* (*NER*) dengan metode *hybrid Bi-LSTM* dan *Transformer*?
2. Apakah penerapan metode *Bi-LSTM* dan *Transformer* dapat meningkatkan performa pengenalan entitas pada teks berita online Bahasa Indonesia menggunakan *Named Entity Recognitin (NER)*?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut :

1. Menganalis metode ekstraksi entitas *Named Entity Recognition (NER)* pada berita online Bahasa Indonesia.

2. Meningkatkan akurasi pengenalan entitas dalam sebuah teks berita online Bahasa Indonesia dengan penerapan metode *Bi-LSTM* dan *Transformer*.

1.4 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut :

1. Memberikan kontribusi pada pengembangan *Named Entity Recognition (NER)* untuk Bahasa Indonesia yang dapat digunakan untuk sistem *Natural Language Processing (NLP*) lainnya.
2. Sebagai referensi atau rujukan baru dalam dunia akedimisi khususnya pada proses pengembangan *Named Entity Recognition* (NER) untuk Bahasa Indonesia.

1.5 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah :

1. Pada penelitian ini fokus utama penulis adalah pengembangan pada tugas *named entity recognition (NER).*
2. Data yang diolah adalah berita berbahasa Indonesia.
3. Entitas yang akan diekstrak adalah entitas person, organization, location, quantity, time dan outside.

BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka

2.1.1 *Natural Language Processing*

Natural language processing (pemrosesan/ pengolahan bahasa alami) adalah metode yang memproses input teks menjadi kata-kata kunci jawaban user. (Hartanto dkk, 2013:35). NLP merupakan kombinasi dari ilmu komputer dan bidang kecerdasan buatan yang berkaitan dengan linguistik. Selain berhubungan dengan lingistik NLP juga berkaitan dengan bagaimana mesin dapat memahami bahasa manusia agar keduanya dapat berinteraksi. Dengan adanya NLP, komputer dapat belajar dan memahami bahasa manusia, sehingga mesin dapat berkomunikasi dengan manusia. Dalam hal ini agar suatu komputer memahami bahasa alami, ia harus memiliki pengetahuan tentang bahasa alami itu sendiri baik dari segi kata yang digunakan, arti dari kata tersebut, fungsi kata dari sebuah kalimat dan bagaimana dari kata-kata tersebut dapat membentuk sebuah kalimat. NLP menggabungkan komputasi linguistik pemodelan bahasa manusia berbasis aturan dengan model statistik, pembelajaran mesin, dan pembelajaran mendalam. Bersama-sama, teknologi ini memungkinkan komputer untuk memproses bahasa manusia dalam bentuk teks atau data suara dan untuk memahami maknanya sepenuhnya, lengkap dengan maksud dan sentimen pembicara atau penulis.

Sampai saat ini penggunaan NLP sudah sangat berkembang. Secara tidak langsung kemungkinan anda telah berinteraksi dengan NLP, namun dalam bentuk yang lebih kompleks dan berupa software. Penerapan NLP dalam kehidupan sehari-hari bisa anda lihat pada aplikasi pencarian seperti Google, Alat penerjemah bahasa, sistem GPS yang dapat dioperasikan dengan suara, perangkat lunak yang digunakan untuk mendikte ucapan ke teks, dan chatbot pada pusat layanan pelanggan. NLP memiliki beberapa pembagian tugas dalam melakukan pemecahan data teks dan suara manusia untuk membantu komputer memahami inputan bahasa manusia. Beberapa tugas tersebut adalah sebagai berikut :

a. Speech Recognition (Pengenalan Suara)

Speech Recognition atau yang biasa dikenal dengan automatic speech recognition (ASR) merupakan suatu pengembangan teknik dan sistem yang memungkinkan komputer untuk menerima masukan berupa kata yang diucapkan. Teknologi ini memungkinkan suatu perangkat untuk mengenali dan memahami kata-kata yang diucapkan dengan cara digitalisasi kata dan mencocokkan sinyal digital tersebut dengan suatu pola tertentu yang tersimpan dalam suatu perangkat. Kata-kata yang diucapkan diubah bentuknya menjadi sinyal digital dengan cara mengubah gelombang suara menjadi sekumpulan angka yang kemudian disesuaikan dengan kode-kode tertentu untuk mengidentifikasikan kata-kata tersebut. Hasil dari identifikasi kata yang diucapkan dapat ditampilkan dalam bentuk tulisan atau dapat dibaca oleh perangkat teknologi sebagai sebuah komando untuk melakukan suatu pekerjaan, misalnya penekanan tombol pada telepon genggam yang dilakukan secara otomatis dengan komando suara.

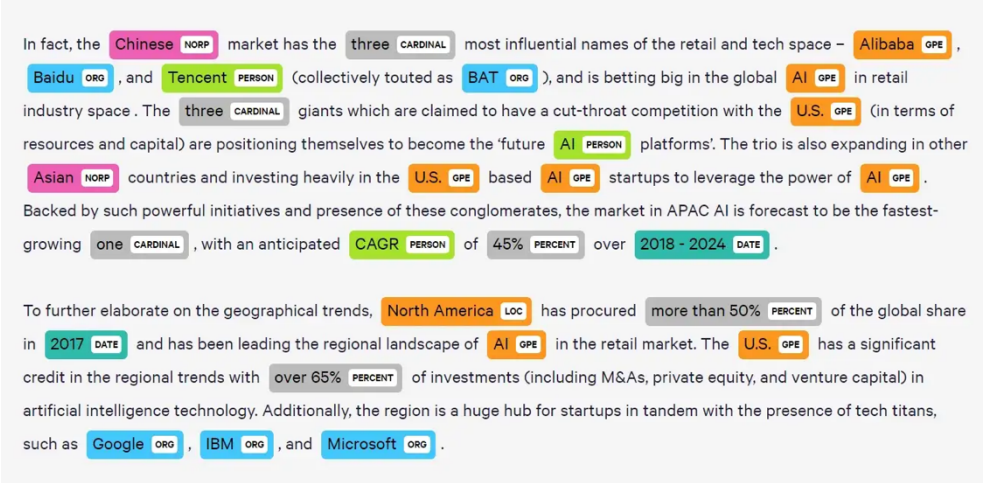
b. Sentiment Analysis (Analisis Sentimen)

Menurut Liu (2008), sentiment analysis (analisis sentimen) atau sering disebut juga dengan opinion mining (penambangan opini) adalah studi komputasi untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, subjektifitas, penilaian atau pandangan yang terdapat dalam suatu teks. Dave et al (2003), menjelaskan bahwa sebuah alat bantu penambangan opini merupakan pemrosesan sekumpulan hasil pencarian dari suatu item yang diberikan, menghasilkan satu daftar atribut produk (misal kualitas, fitur, dan lain-lain) dan menghitung agregasi dari opini dari masing-masing atribut tersebut (rendah, sedang, tinggi).

c. Named Entity Recognition (Pengenalan Entitas Bernama)

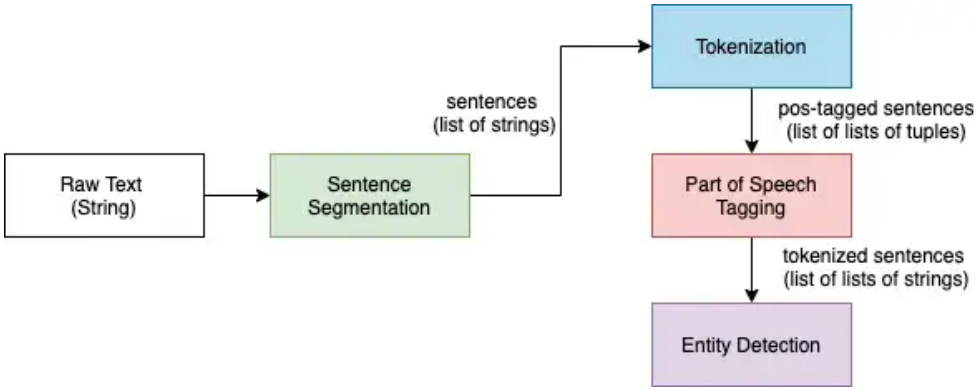
Named Entity Recognition adalah sub-bidang dalam pemrosesan bahasa alami (Natural Language Processing) dan salah satu komponen utama dari information extraction yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas yang tercantum pada suatu teks. Tujuan yang diharapkan dari proses dalam NER adalah untuk melakukan ekstraksi dan klasifikasi nama ke dalam beberapa kategori

dengan mengacu kepada makna yang tepat (Mansouri, et al., 2008). Named Entity Recognition umumnya digunakan untuk mendeteksi entitas berupa nama orang, lokasi, organisasi, tanggal, dan nomor telepon, serta entitas lain yang memiliki nama atau label khususe. Proses NER biasanya dimulai dengan memisahkan token-token dalam teks menjadi entitas yang terpisah, seperti kata, frasa, atau klausa. Kemudian, setiap entitas tersebut diklasifikasikan berdasarkan jenis entitas yang sesuai, misalnya sebagai nama orang, lokasi, atau organisasi. Pada akhirnya, entitas yang telah dikenali akan ditandai dengan label atau tag khusus yang menunjukkan jenis entitas tersebut.



Gambar 1. Keluaran Ekstrajsi Entitas NER

NER merupakan salah satu teknik yang sering digunakan dalam bidang Natural Language Processing (NLP) atau pengolahan bahasa alami. Ada beberapa cara yang dapat dilakukan untuk melakukan NER, salah satunya adalah dengan menggunakan rule-based approach. Pada pendekatan ini, kita menentukan aturan-aturan yang akan digunakan untuk mengidentifikasi entitas bernama dalam teks. Aturan-aturan tersebut dapat berupa kata-kata atau karakter khusus yang biasanya muncul di sekitar entitas bernama, seperti tanda koma atau titik di sebelah nama orang. Selain itu, NER juga dapat dilakukan dengan menggunakan supervised learning. Pada pendekatan ini, kita memberikan contoh-contoh teks yang telah dilabeli dengan entitas bernama yang terdapat di dalamnya. Kemudian, kita entrain sebuah model untuk dapat mengenali entitas bernama dalam teks yang belum diketahui.



Gambar 2 Proses ekstraksi entitas

Model tersebut dapat berupa model supervised learning biasa, seperti model support vector machine (SVM) atau model Naive Bayes, atau model deep learning, seperti model convolutional neural network (CNN) atau model recurrent neural network (RNN). Setelah model tersebut terentrain, kita dapat menggunakannya untuk mengidentifikasi entitas bernama dalam teks yang belum diketahui. Proses ini biasanya dilakukan dengan memecah teks menjadi token-token kemudian menganalisis masing-masing token untuk menentukan apakah token tersebut merupakan entitas bernama atau bukan. Selain itu, NER juga dapat dilakukan dengan menggunakan teknik yang disebut transfer learning. Pada pendekatan ini, kita menggunakan sebuah model yang telah terentrain pada dataset yang besar dan banyak, kemudian kita fine-tune model tersebut untuk menyesuaikannya dengan dataset yang akan kita gunakan. Dengan cara ini, kita dapat mempercepat proses pelatihan model karena model tersebut telah memiliki kapasitas yang cukup untuk mengenali entitas bernama dalam teks.

Terdapat beberapa kelebihan yang dimiliki oleh NER dibandingkan dengan teknik pengolahan teks lainnya:

a. Pertama, NER memiliki kemampuan untuk mengidentifikasi entitas bernama dengan tingkat akurasi yang tinggi. Ini karena NER biasanya menggunakan model deep learning yang telah terentrain pada dataset yang besar dan banyak, sehingga dapat mengenali entitas bernama dengan tingkat akurasi yang tinggi.
b. Kedua, NER memiliki kemampuan untuk mengidentifikasi entitas bernama yang tidak terstruktur. Ini berarti bahwa NER dapat digunakan untuk mengidentifikasi entitas bernama dalam teks yang tidak memiliki struktur yang

teratur, seperti teks berita atau teks obrolan. Dengan cara ini, kita dapat mengambil informasi penting dari teks yang tidak terstruktur dengan mudah.

c. Ketiga, NER memiliki kemampuan untuk mengidentifikasi entitas bernama dalam bahasa yang berbeda. Ini karena NER biasanya menggunakan model deep learning yang dapat mempelajari bahasa secara umum, sehingga dapat digunakan untuk mengidentifikasi entitas bernama dalam bahasa apapun. Dengan cara ini, kita dapat mengambil informasi penting dari teks yang ditulis dalam bahasa yang berbeda dengan mudah.

Secara keseluruhan, NER merupakan teknik yang berguna dalam bidang NLP karena memiliki kemampuan untuk mengidentifikasi entitas bernama dengan tingkat akurasi yang tinggi, dapat digunakan untuk mengidentifikasi entitas bernama yang tidak terstruktur, dan dapat digunakan untuk mengidentifikasi entitas bernama dalam bahasa yang berbeda. Dengan menggunakan NER, kita dapat mengambil informasi penting dari teks dengan mudah dan efisien.

2.1.2 *Information extraction (IE)*

Information extraction (IE) adalah teknik pemrosesan bahasa alami (natural language processing) yang bertujuan untuk mengekstraksi informasi yang tersimpan dalam teks dan mengklasifikasikannya ke dalam kategori yang terstruktur. Proses IE biasanya dimulai dengan mengidentifikasi entitas-entitas yang tercantum dalam teks, seperti nama orang, lokasi, organisasi, dan tanggal, menggunakan teknik seperti named entity recognition (NER). Kemudian, entitas-entitas tersebut diklasifikasikan ke dalam kategori yang sesuai, seperti nama orang, lokasi, atau organisasi, dan ditandai dengan label atau tag khusus. Information extraction banyak digunakan dalam berbagai aplikasi, seperti pemrosesan teks, pemahaman mesin, dan penelitian. Misalnya, IE dapat digunakan untuk mengekstraksi informasi tentang nama orang dalam sebuah artikel berita dan menggunakannya untuk membuat indeks berita yang terorganisir. IE juga dapat digunakan untuk mengklasifikasikan teks berita berdasarkan lokasi yang tercantum di dalamnya dan menampilkannya dalam peta agar pengguna dapat dengan mudah menemukan lokasi yang dimaksud.

Ada beberapa metode yang dapat digunakan untuk melakukan information extraction, di antaranya adalah metode statistik, metode rule-based, dan metode deep learning. Metode statistik menggunakan algoritma yang melakukan klasifikasi berdasarkan pola dan frekuensi yang terdapat dalam teks, sementara metode rule-based menggunakan aturan-aturan yang telah ditetapkan sebelumnya untuk mengklasifikasikan entitas dalam teks. Metode deep learning, di sisi lain, menggunakan jaringan saraf tiruan untuk belajar secara otomatis cara mengklasifikasikan entitas dalam teks. Information extraction banyak digunakan dalam berbagai aplikasi, seperti pemrosesan teks, pemahaman mesin, dan penelitian. Misalnya, IE dapat digunakan untuk mengklasifikasikan nama orang dalam sebuah artikel berita dan menggunakannya untuk membuat indeks berita yang terorganisir. IE juga dapat digunakan untuk mengekstraksi informasi tentang lokasi dalam sebuah teks berita dan menampilkannya dalam peta untuk memudahkan pengguna menemukan lokasi yang dimaksud.

2.1.3 *Text Mining*

Text mining merupakan suatu proses untuk mengekstrak pola dalam mengeksplorasi pengetahuan dari sumber data yang berbentuk teks. Proses Text mining dimulai dengan mengumpulkan data dari berbagai sumber yang tersedia dalam berbagai format file seperti teks biasa, halaman web, file pdf dan sebagainya. Kemudian melakukan pre-processing dan pembersihan data dilakukan untuk mendeteksi dan menghapus anomali pada data. Proses pembersiha harus memastikan untuk menangkap esensi teks sebenarnya yang tersedia. Pemrosesan dan pengedalian diterapkan untuk mengaudit kemudian membersihkan data dengan pemrosesan otomatis. Setelah itu dilakukan analisis pola pada data guna memperoleh informasi yang berharga dan relevan (Talib, et al., 2016).

Tujuan text mining untuk menggali informasi yang dapat berguna dari beberapa Dokumen, Selain itu text mining dapat mendukung proses knowledge discovery pada beberapa Dokumen yang besar. Berdasarkan tujuannya tersebut text mining dapat diterapkan ke berbagai bidang yang cukup luas. Terdapat beberapa area penerapa text mining, yaitu: ekstaksi informasi (ekstaksi informasi), pelacakan topik (topic tracking), perangkuman (Summarization), kategorisasi

(Categorization), penggugusan (clustering), penautan konsep (concept linking), dan penjawaban pertanyaan (question answering).

2.1.4 *Part of Speech (PoS) Tagging*

Part-of-speech (POS) tagging atau secara singkat dapat ditulis sebagai tagging merupakan proses pemberian penanda POS atau kelas sintaktik pada tiap kata di dalam corpus. Dikarenakan tag secara umum juga diaplikasikan pada tanda baca, maka dalam proses tagging, tanda baca seperti tanda titik, tanda koma, dll perlu dipisahkan dari kata-kata. Oleh sebab itu, proses tokenisasi biasanya dilakukan sebelum POS tagging. Selain itu beberapa preprocessing juga dilakukan seperti pemisahan koma, tanda petik, dll dari kata serta dilakukan juga disambiguitas pada tanda baca penanda akhir kalimat seperti tanda titik dan tanda tanya agar dapat dibedakan dari tanda yang digunakan untuk singkatan (seperti contohnya: e.g. dan etc.). Seperti pernah disebutkan sebelumnya, masalah utama dalam melakukan tagging adalah ambiguitas terutama ketika kita meminta sistem untuk melakukannya secara otomatis. Contoh dari beberapa kata yang seringkali menimbullkan ambiguitas diantaranya adalah book dikarenakan memiliki 2 buah makna, yakni book sebagai kata benda yang berarti buku dan sebagai kata kerja yang berarti memesan. Oleh karena itu POS-tagging bertujuan untuk menyelesaikan masalah ini dengan memilih tag yang tepat untuk konteks kata di dalam kalimat.

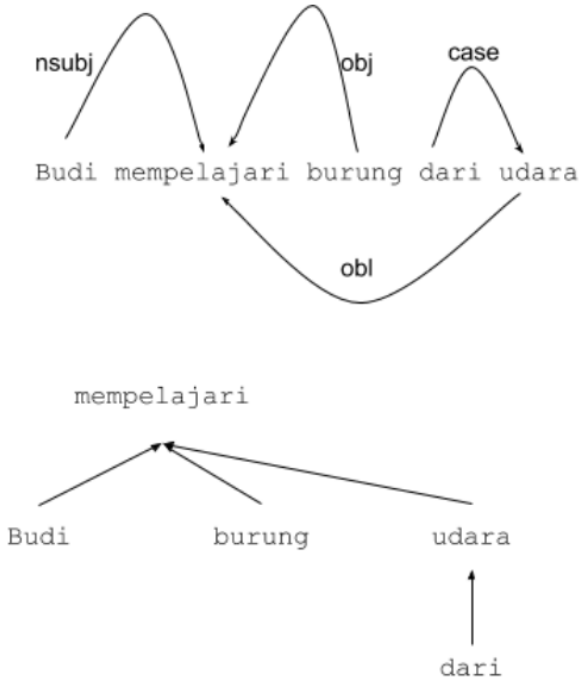
Kebanyakan algoritma untuk tagging termasuk salah satu kelas dari rule-based taggers dan stochastic taggers. Rule-based tagger secara umum melibatkan database dalam ukuran yang besar mengenai aturan-aturan disambiguasi dari tulisan tangan yang menspesifikasikan diantaranya, sebuah kata yang ambigu adalah kata benda dan bukan kata kerja jika diikuti oleh determiner. Salah satu contoh rule-based tagger adalah EngCG, yang berdasarkan arsitektur Constraint Grammar dari Karlsson et al (1995). Stochastic taggers secara umum menyelesaikan masalah ambiguitas pada tagging dengan menggunakan korpus yang dilatih untuk menghitung probabilitas dari sebuah kata yang dengan tag yang diberikan dalam sebuah konteks.

Part of speech (PoS) dapat digunakan sebagai salah satu faktor dalam pemrosesan NER. Misalnya, sebuah kata yang memiliki part of speech sebagai noun (kata benda) kemungkinan besar merupakan entitas bernama. Sebagai contoh, kalimat "John membeli sebuah mobil baru" mengandung entitas bernama "John" dan "mobil" yang memiliki part of speech sebagai noun. Selain itu, part of speech juga dapat digunakan untuk membedakan antara entitas bernama yang berbeda. Misalnya, kalimat "John pergi ke New York" mengandung entitas bernama "John" dan "New York" yang memiliki part of speech sebagai noun. Dengan menggunakan informasi part of speech, kita dapat mengetahui bahwa "John" dan "New York" adalah entitas bernama yang berbeda. Namun, perlu diingat bahwa part of speech tidak selalu merupakan indikator yang pasti untuk mengidentifikasi entitas bernama. Terkadang, ada entitas bernama yang memiliki part of speech yang berbeda, seperti kata ganti ("I", "you", dll.) atau kata seru ("oh", "alas", dll.). Oleh karena itu, NER biasanya menggunakan metode yang lebih kompleks dan mengintegrasikan banyak faktor, termasuk part of speech, untuk mengidentifikasi entitas bernama dengan benar.

2.1.5 *Dependency Parsing*

Dependency parsing adalah suatu teknik dalam natural language processing (NLP) yang digunakan untuk mengidentifikasi hubungan gramatikal antara satu kata dengan kata lain dalam suatu kalimat. Dengan menggunakan dependency parsing, kita dapat memahami struktur gramatikal kalimat dan mengidentifikasi bagaimana setiap kata terhubung dengan kata lain dalam kalimat tersebut. Menurut ahli, dependency parsing merupakan salah satu metode yang efektif untuk menganalisis struktur gramatikal kalimat. Teknik ini memperlihatkan hubungan antara satu kata dengan kata lain secara visual, sehingga memudahkan kita untuk memahami struktur kalimat dan makna yang terkandung di dalamnya. Selain itu, dependency parsing juga dapat digunakan dalam berbagai aplikasi NLP, seperti pemrosesan kalimat, pengenalan entitas bernama, dan pemodelan inten. Dengan menggunakan dependency parsing, kita dapat mengidentifikasi bagaimana setiap kata dalam kalimat terhubung satu sama lain, sehingga dapat membantu kita untuk memahami makna kalimat secara keseluruhan.

Dependency parsing dapat memainkan peran yang penting dalam named entity recognition (NER), yaitu proses pengenalan entitas bernama dalam teks. Dengan menggunakan dependency parsing, kita dapat mengidentifikasi hubungan gramatikal antara satu kata dengan kata lain dalam suatu kalimat, sehingga memudahkan kita untuk mengidentifikasi entitas bernama dalam teks. Misalnya,



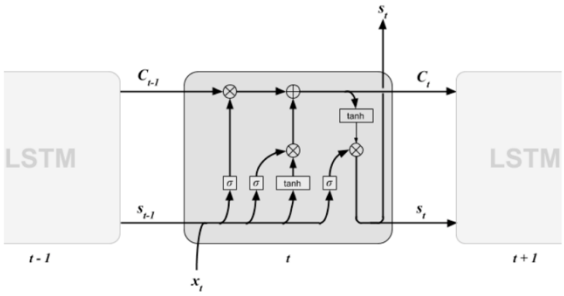
Gambar 3 Skema *dependency parsing*

dalam kalimat "John membeli sebuah mobil baru", kata "John" dan "mobil" memiliki hubungan gramatikal sebagai subjek dan objek. Dengan menggunakan dependency parsing, kita dapat mengidentifikasi bahwa "John" dan "mobil" merupakan entitas bernama yang berbeda. Selain itu, dependency parsing juga dapat digunakan untuk mengidentifikasi entitas bernama yang memiliki bentuk yang tidak biasa, seperti kata ganti atau kata seru. Misalnya, dalam kalimat "Saya pergi ke New York", kata "Saya" memiliki hubungan gramatikal sebagai subjek. Dengan menggunakan dependency parsing, kita dapat mengidentifikasi bahwa "Saya" merupakan entitas bernama yang berbeda dari "New York". Oleh karena itu, dependency parsing dapat membantu kita untuk mengidentifikasi entitas bernama

dalam teks dengan lebih akurat dan memahami struktur gramatikal kalimat secara keseluruhan. Namun, perlu diingat bahwa dependency parsing hanya merupakan salah satu metode yang dapat digunakan dalam NER, dan biasanya digunakan bersama dengan metode lain untuk meningkatkan akurasi pengenalan entitas bernama.

2.1.6 Bidirectional Long Short-Term Memory

Bi-directional merupakan metode pengembangan dari metode sebelumnya yaitu LSTM. Pada dasarnya, LSTM terdiri dari tiga gerbang perkalian (multiplicative gates) yang mengontrol jumlah informasi yang akan dihilangkan, diteruskan, dan luaran langkah waktu berikutnya (t + 1). Gambar 4 berikut ini adalah ilustrasi sederhana dari LSTM.



Gambar 4 Skema LSTM

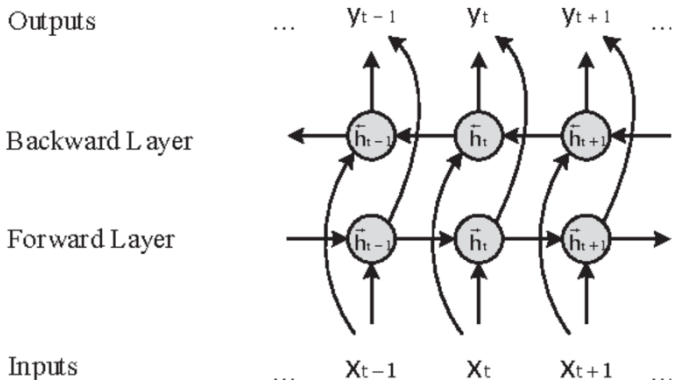
Fitur utama dari LSTM adalah adanya sebuah jalur yang menghubungkan cell state sebelumnya (C_{t-1}) ke cell state waktu sekarang (C_t). Dengan adanya jalur tersebut, suatu nilai di C_{t-1} dengan mudah diteruskan ke C_t dengan sedikit sekali modifikasi, sehingga dapat menghindari terjadinya vanishing dan exploding gradient.

Fitur lain dari LSTM adalah adanya gerbang sigmoid yang mengatur seberapa banyak informasi yang dapat dilewatkan dari data masukan. Untuk suatu masukan vektor x, luaran dari gerbang sigmoid adalah ($A \cdot x + b$), dimana A adalah nilai bobot, b adalah bias, keduanya dipelajari selama proses pelatihan, dan σ adalah fungsi sigmoid. Luaran gerbang sigmoid tersebut digunakan untuk mengontrol seberapa banyak informasi dari C_{t-1} yang akan diteruskan ke C_t. Luaran gerbang

sigmoid tersebut adalah angka antara 0 dan 1; luaran 0 artinya tidak ada informasi yang dilewatkan, sedangkan 1 artinya seluruh informasi dilewatkan.

Gerbang sigmoid pertama dari kiri yang ditunjukkan pada Gambar 2.3 disebut dengan forget gate (f_t). Pada gerbang sigmoid tersebut akan diputuskan informasi apa yang akan dilupakan dari ' masukan C_{t-1}. Masukan dari gerbang tersebut adalah nilai S_{t-1} dan X_t, dan menghasilkan nilai angka antara 0 dan 1 untuk setiap elemen dalam C_{t-1} Gerbang selanjutnya adalah gerbang yang digunakan untuk mengontrol informasi baru apa yang digunakan di C_t yang disebut dengan input gate (i_t). Gerbang sigmoid i_t memutuskan nilai mana yang akan diperbarui, kemudian sebuah lapisan $tanh$ menghasilkan kandidat vektor cell state baru C_t yang selanjutnya digabungkan untuk membuat pembaruan ke cell state C_t.

Cell state yang baru (C_t) diperoh dengan mengalikan cell state lama (C_{t-1}) dengan nilai forget gate (f_t) dan kemudian ditambahkan dengan hasil perkalian C_t dan input gate (i_t). Gerbang terakhir adalah output gate (o_t) yang merupakan gerbang sigmoid paling kanan pada Gambar 2.3 untuk memutuskan bagian-bagian C_t yang akan dihasilkan. Kemudian, nilai C_t dilewatkan pada fungsi aktivasi $tanh$ untuk membuat nilainya menjadi antara -1 dan 1, dan dikalikan dengan o_t sehingga hanya dihasilkan bagian yang telah diputuskan.



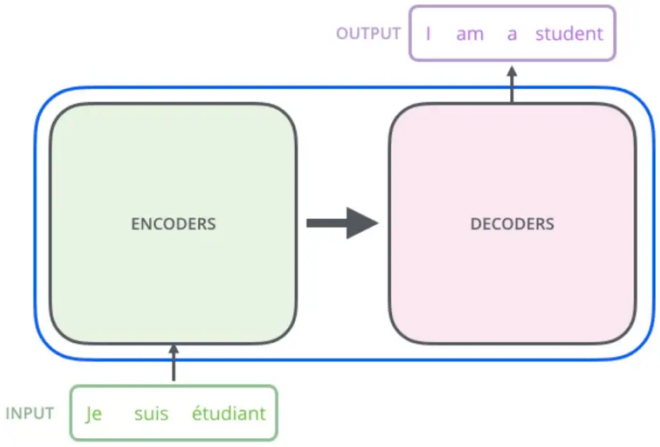
Gambar 5 *Bidirectional LSTM*

Untuk pelabelan sekuensial seperti NER dan speech recognition, LSTM sangat bermanfaat apabila ada akses untuk mengetahui informasi waktu sebelum dan setelahnya [26]. Namun, vektor luaran pada LSTM (S_t) hanya menerima informasi dari waktu sebelumnya saja, tidak punya akses informasi untuk waktu setelahnya. Solusi permasalahan tersebut terselesaikan berkat ditemukannya model

LSTM yang memiliki akses informasi dua arah yang disebut dengan bi-directional LSTM (BiLSTM) [27]. Arsitektur model bidirectional LSTM ditunjukkan pada Gambar 5. Ide dasarnya adalah menggunakan forward dan backward LSTM untuk menangkap informasi waktu sebelum dan setelahnya. Kemudian luaran dari forward dan backward LSTM tersebut digabungkan untuk memperoleh nilai luaran akhir.

2.1.7 *Transformer*

Transformer adalah sebuah arsitektur deep learning yang digunakan untuk pengolahan teks. Arsitektur ini pertama kali diperkenalkan oleh Vaswani et al. dalam paper mereka yang berjudul "Attention is All You Need" pada tahun 2017. *Transformer* memiliki beberapa kelebihan dibandingkan dengan arsitektur deep learning lainnya, sehingga cukup populer digunakan dalam bidang pengolahan teks. Menurut ahli, salah satu kelebihan utama dari *Transformer* adalah kemampuannya untuk melakukan paralelisasi yang lebih baik dibandingkan dengan arsitektur deep learning lainnya. Pada arsitektur deep learning lainnya, seperti recurrent neural network (RNN) atau convolutional neural network (CNN), sebuah unit pengolahan harus menunggu hasil dari unit pengolahan sebelumnya sebelum dapat diproses. Dengan cara ini, arsitektur deep learning lainnya tidak dapat dilakukan secara paralel. Namun, pada *Transformer*, unit-unit pengolahan tidak saling tergantung satu sama lain. Dengan cara ini, semua unit pengolahan dapat diproses secara



Gambar 6 Proses kerja *Transformer*

paralel, sehingga *Transformer* dapat melakukan pengolahan teks dengan lebih cepat

dibandingkan dengan arsitektur deep learning lainnya. Secara umum *Transformer* terdiri atas dua bagian utama, yaitu encoder dan decoder.

Selain itu, *Transformer* juga memiliki kemampuan untuk melakukan generalisasi yang lebih baik dibandingkan dengan arsitektur deep learning lainnya. Ini karena *Transformer* menggunakan teknik self-attention untuk mengolah teks, dimana setiap unit pengolahan dapat memperhatikan semua unit pengolahan lain dalam mengambil keputusan. Dengan cara ini, *Transformer* dapat menangkap struktur dan pola yang lebih abstrak dalam data, sehingga dapat melakukan generalisasi dengan lebih baik. *Transformer* juga memiliki beberapa kekurangan. Pertama, *Transformer* membutuhkan jumlah data yang besar untuk pelatihan. Hal ini karena *Transformer* menggunakan teknik self-attention yang membutuhkan banyak parameter untuk pelatihan. Jika data yang tersedia tidak cukup besar, maka *Transformer* akan kesulitan untuk melakukan pelatihan dan menghasilkan hasil yang baik. Kedua, *Transformer* juga membutuhkan waktu dan komputasi yang cukup banyak untuk pelatihan. Hal ini karena *Transformer* menggunakan banyak parameter dan membutuhkan banyak data untuk pelatihan, sehingga membutuhkan waktu dan komputasi yang cukup banyak untuk melakukan pelatihan. Jika kita tidak memiliki waktu dan komputasi yang cukup, maka *Transformer* tidak akan dapat melakukan pelatihan dengan baik. Namun, meskipun memiliki kekurangan, *Transformer* tetap merupakan arsitektur deep learning yang berguna dalam bidang pengolahan teks karena dapat melakukan paralelisasi yang lebih baik, dapat melakukan generalisasi yang lebih baik, dan dapat menangkap struktur dan pola yang lebih abstrak dalam data. Dengan menggunakan *Transformer*, kita dapat melakukan pengolahan teks dengan lebih cepat dan lebih akurat dibandingkan dengan arsitektur deep learning lainnya

Pada *Transformer* terdapat penggunaan teknik yang disebut self-attention. Self-attention adalah sebuah teknik yang digunakan dalam model deep learning untuk memproses input secara paralel dan menghasilkan output yang lebih akurat daripada model-model sebelumnya seperti recurrent neural networks (RNNs). Self-attention bekerja dengan menghitung bobot untuk setiap pasangan elemen dalam input dan menggunakan bobot tersebut untuk mengkombinasikan elemen-elemen tersebut menjadi output. Dalam model deep learning, self-attention dapat

diterapkan pada sebuah layer yang disebut "self-attention layer". Layer ini menerima input vektor-vektor dan menghitung bobot self-attention untuk setiap pasangan elemen dalam input. Setelah itu, bobot tersebut digunakan untuk mengkombinasikan elemen-elemen input menjadi sebuah output vektor baru yang mengandung informasi yang relevan dari input.

Transformer dapat digunakan dalam NER (Named Entity Recognition) untuk mengidentifikasi dan menandai entitas-entitas yang terdapat dalam sebuah teks, seperti nama orang, tempat, atau perusahaan. *Transformer* dapat melakukan NER dengan lebih baik dibandingkan dengan metode NER lainnya karena dapat melakukan paralelisasi yang lebih baik, dapat melakukan generalisasi yang lebih baik, dan dapat menangkap struktur dan pola yang lebih abstrak dalam data.

Menurut ahli, salah satu kelebihan utama dari *Transformer* dalam NER adalah kemampuan untuk melakukan pemodelan konteks secara efektif. Pada metode NER lainnya, seperti rule-based atau CRF (Conditional Random Fields), entitas yang dikenali hanya tergantung pada fitur-fitur yang ditentukan secara manual. Namun, pada *Transformer*, entitas yang dikenali juga tergantung pada konteks dari entitas tersebut dalam teks. Dengan cara ini, *Transformer* dapat menangkap pola yang lebih rumit dalam data dan mengidentifikasi entitas dengan lebih akurat. Selain itu, *Transformer* juga dapat menangkap hubungan antar entitas yang lebih rumit dibandingkan dengan metode NER lainnya. Pada metode NER lainnya, hubungan antar entitas hanya dapat ditentukan secara manual dengan menggunakan aturan-aturan yang dibuat oleh ahli. Namun, pada *Transformer*, hubungan antar entitas dapat ditentukan secara otomatis dengan menggunakan teknik self-attention yang dimiliki oleh *Transformer*. Dengan cara ini, *Transformer* dapat mengidentifikasi entitas dan hubungan antar entitas dengan lebih akurat.

2.1.8 Embedding Text

Pada tahun 2003, Bengio dkk. (Bengio et al., 2003) memperkenalkan istilah word embedding. Word embedding merupakan sebuah representasi vektor untuk kata-kata dalam sebuah teks. Tujuan dari word embedding ini adalah untuk mengkode kata-kata dalam teks ke dalam bentuk vektor numerik yang dapat dimengerti dan diproses oleh mesin. Dengan cara ini, kita dapat melakukan analisis

terhadap teks dengan menggunakan metode-metode statistik atau deep learning. Keunggulan word embedding tidak membutuhkan anotasi, dapat langsung diturunkan dari korpus tak teranotasi. Word embedding dapat dibuat langsung dari dataset yang dimiliki atau menggunakan pre-trained word embedding yang telah tersedia. Pre-trained word embedding ini adalah word embedding yang telah dilatih menggunakan dataset yang besar pada domain permasalahan tertentu yang dapat digunakan untuk menyelesaikan permasalahan lain yang serupa.

Menurut ahli, word embedding memiliki beberapa kelebihan dibandingkan dengan representasi kata-kata yang lain. Pertama, word embedding memiliki ukuran yang lebih kecil dibandingkan dengan representasi kata-kata seperti one-hot encoding, sehingga dapat mengurangi overhead komputasi dalam pelatihan model deep learning. Kedua, word embedding memiliki struktur yang lebih baik dibandingkan dengan representasi kata-kata seperti one-hot encoding. Karena kata-kata yang mirip akan memiliki vektor yang juga mirip dalam word embedding, maka model deep learning yang menggunakan word embedding akan lebih mudah untuk mempelajari struktur data. Ketiga, word embedding memiliki interpretabilitas yang lebih baik dibandingkan dengan representasi kata-kata seperti one-hot encoding. Karena vektor dalam word embedding dapat diturunkan dari data, maka kita dapat melihat hubungan antara kata-kata dengan melihat jarak antara vektor-vektor yang merepresentasikannya.

Untuk menghasilkan word embedding, kita dapat menggunakan beberapa cara, salah satunya adalah dengan menggunakan model word2vec. Model ini merupakan sebuah model deep learning yang dapat mempelajari word embedding dari data teks yang besar dan banyak. Model ini biasanya dibangun dengan menggunakan jaringan neural network yang terdiri dari dua buah bagian, yaitu bagian encoder dan bagian decoder. Bagian encoder akan memproses teks dan mengkode setiap kata ke dalam bentuk vektor numerik. Kemudian, bagian decoder akan memproses vektor-vektor tersebut dan menghasilkan prediksi untuk kata-kata yang seharusnya muncul di sekitar kata yang sedang diproses. Dengan cara ini, model akan belajar untuk menghasilkan word embedding yang terrepresentasi dengan baik dari data teks. Setelah model terlatih, kita dapat menggunakannya untuk mengkode kata-kata dalam teks baru menjadi vektor-vektor numerik yang

dapat digunakan untuk keperluan analisis teks. Selain itu, kita juga dapat menggunakan word embedding yang dihasilkan oleh model ini sebagai input untuk model deep learning lainnya, seperti model SVM atau model Naive Bayes, untuk melakukan analisis teks yang lebih lanjut. Secara keseluruhan, word embedding merupakan sebuah representasi yang berguna untuk kata-kata dalam teks karena memiliki ukuran yang kecil, struktur yang baik, dan interpretabilitas yang baik. Dengan menggunakan word embedding, kita dapat melakukan analisis teks dengan lebih efisien dan lebih mudah dengan menggunakan metode-metode statistik atau deep learning.

Selain model word2vec, ada juga beberapa model lain yang dapat digunakan untuk menghasilkan word embedding, seperti model GloVe atau model fastText. Model-model ini memiliki cara kerja yang mirip dengan model word2vec, namun memiliki beberapa kelebihan dan kekurangan masing-masing. Contohnya, model GloVe menggunakan teknik yang disebut factorization untuk menghasilkan word embedding, sehingga dapat menghasilkan word embedding yang lebih akurat dibandingkan dengan model word2vec. Namun, model GloVe membutuhkan lebih banyak waktu dan komputasi untuk menghasilkan word embedding dibandingkan dengan model word2vec. Sedangkan model fastText menggunakan teknik yang disebut bag-of-ngrams untuk menghasilkan word embedding. Dengan cara ini, model fastText dapat menghasilkan word embedding untuk kata-kata yang belum pernah muncul dalam data latih, sehingga dapat menangani masalah out-of-vocabulary (OOV) yang sering muncul dalam pengolahan teks. Namun, model fastText cenderung menghasilkan word embedding yang kurang akurat dibandingkan dengan model GloVe atau model word2vec. Dalam memilih model untuk menghasilkan word embedding, kita harus mempertimbangkan kebutuhan dan karakteristik data yang akan kita analisis. Jika kita membutuhkan word embedding yang akurat dan tidak terlalu peduli dengan waktu dan komputasi yang dibutuhkan, maka model GloVe dapat menjadi pilihan yang baik. Namun, jika kita membutuhkan word embedding yang dapat menangani masalah OOV dan tidak terlalu peduli dengan akurasi, maka model fastText dapat menjadi pilihan yang baik.

2.2 Metode Penyelesaian Masalah

2.2.1 Metode Pengenalan Entitas Bernama

Pengenalan entitas bernama dengan pendekatan metode deep learning bagian NLP menggunakan pemodelan *Transformer*. *Transformer* adalah sebuah arsitektur deep learning yang digunakan untuk pengolahan teks. salah satu kelebihan utama dari *Transformer* adalah kemampuannya untuk melakukan paralelisasi yang lebih baik dibandingkan dengan arsitektur deep learning lainnya. Pada arsitektur deep learning lainnya, seperti recurrent neural network (RNN) atau convolutional neural network (CNN), sebuah unit pengolahan harus menunggu hasil dari unit pengolahan sebelumnya sebelum dapat diproses. Dengan cara ini, arsitektur deep learning lainnya tidak dapat dilakukan secara paralel. Namun, pada *Transformer*, unit-unit pengolahan tidak saling tergantung satu sama lain sehingga unit pengolahan lain dapat diproses secara paralel dan berdampak pada waktu yang dibutuhkan untuk melakukan pengenalan entitas bernama.

Bagian dari metode ini yang akan penulis modifikasi adalah pada pagian preprocessing NLP. Pada bagian preprocessing pengenalan entitas bernama terdapat beberapa tahapan yaitu cleaning, case folding, stopword, tokenizing, spelling normalisation, filtering, stemming, dan penambahan tag. Bagian yang akan dimodifikasi adalah pada bagian penambahan tag. Selain itu penulis juga akan mencoba mengulik pada bagian pemodelan NER nya sendiri seperti pada tahapan dependencies parsing di *Transformer* agar didapatkan model yang memiliki performa yang lebih baik untuk pengenalan entitas bernama.

2.2.2 Metode Pengujian Sistem

Pada rencana penelitian ini performa dari model untuk melakukan klasifikasi dapat akan diukur dengan menghitung jumlah kelas yang diprediksi dengan benar (true positive), jumlah yang diprediksi bukan termasuk kelsas tersebut dan benar (true negative), dan yang salah prediksi (false positive atau false negative). Terdapat beberapa metrik klasifikasi yang dapat digunakan, yaitu akurasi, precision, recall, dan fscore. Nilai metrik tersebut dapat dihitung dari confusion matrix sebagai berikut:

a. Precision: Precision mengukur seberapa sering NER menemukan entitas yang benar dari total entitas yang ditemukan. Precision dihitung dengan menggunakan rumus:

$$\boldsymbol{Precission} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})}$$

Di mana True Positive adalah jumlah entitas yang ditemukan oleh NER dan benar, sedangkan False Positive adalah jumlah entitas yang ditemukan oleh NER namun salah. Semakin tinggi nilai precision, semakin tinggi kemungkinan bahwa NER telah menemukan entitas yang benar.

b. Recall: Recall mengukur seberapa sering NER menemukan entitas yang benar dari total entitas yang seharusnya ditemukan. Recall dihitung dengan menggunakan rumus:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})}$$

Di mana True Positive adalah jumlah entitas yang ditemukan oleh NER dan benar, sedangkan False Negative adalah jumlah entitas yang seharusnya ditemukan oleh NER namun tidak ditemukan. Semakin tinggi nilai recall, semakin tinggi kemungkinan bahwa NER telah menemukan semua entitas yang seharusnya ditemukan.

c. F1 score: F1 score merupakan rata-rata harmonis dari precision dan recall. F1 score dihitung dengan menggunakan rumus:

$$\boldsymbol{F1\ Score} = \boldsymbol{2} * \frac{\boldsymbol{Precission} * \boldsymbol{Recall}}{(\boldsymbol{Precission} + \boldsymbol{Recall})}$$

Semakin tinggi nilai F1 score, semakin baik performa NER.

2.2.1 State of The Art

Berikut ini State of The Art Penelitian ini yaitu:

Tabel 1 Matriks jurnal penelitian terkait

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
1.	**Judul:** Ekstraksi Entitas Pada Berita Online Bahasa Indonesia Menggunakan Named Entity Recognition (NER)	**Objek:** Teks/ Korpus **Masalah:** Presisi dari penelitian sebelumnya masih perlu ditingkatkan (87,77%) agar lebih akurat dan natural.	- Melakukan ekstraksi entitas dengan menggunakan metode *Transformer* pada berita online bahasa Indonesia - Melakukan pengembangan pada PoS tagging dan metode *Transformer* pada NER itu sendiri	Penulis melakukan penelitian ini untuk meningkatkan akurasi dan presisi dari hasil kerja Named Entity Recognition agar performa ekstraksi entitas menggunakan NER dapat meningkat
2.	Low Complexity Named-Entity Recognition for Indonesian Language using BiLSTM-CNNs **Nama:** Sahrul Sukardi, Ade Irawan, Meredita Susanti, Randi Fermana Putra **Tahun:** 2020	**Objek:** Teks/Chatbot **Masalah:** NER telah digunakan di banyak bidang pekerjaan salah satunya pada pengembangan chatbot. Penggunaan NLP	BiLSTM-CNNs	Named-entities (NEs) yang digunakan pada penelitian ini hanya terbatas pada nama orang, organisasi, lokasi, kuantitas, dan waktu dengan menggunakan format pelabelan BILOU. Performa model yang dibangun diukur dengan menggunakan

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Penerbit: IEEE	dan deep learning (ML) pada chatbot dapat membuat chatbot lebih cerdas dengan analisis personal yang lebih baik ke pengguna. Tujuan dari penelitian ini adalah untuk membuat model NER dalam bahasa Indonesia dengan menggunakan arsitektur model Bidirectional Long-Short-Term-Memory (BiLSTM) dan Convolutional Neural Networks (CNNs)		metrik evaluasi f1 score dengan rataan mikro. Model BiLSTM-CNNs + pretrained word2vec embedding menghasilkan performa yang sangat baik dibanding model lainnya dengan nilai f1 score rataan mikro 73.37%.
3.	Name Indexing in Indonesian Translation of Hadith using Named Entity Recognition with Naïve Bayes Classifier **Nama:** Fadhila Yasmine Azaliaa, Moch Arif Bijaksanaa, Arief Fatchul Huda	**Objek:** Teks terjemahan Bahsa Indonesia dari Hadist **Masalah:** Banyaknya literatur hadist terkadang menemui masalah untuk mendapatkan informasi yang	NER dengan metode Naïve Bayes Classifier + PoS	Berdasarkan hasil percobaan dengan fitur di NER menggunakan Näıve Bayes, F1-Score tertinggi mencapai 82,63% dengan kombinasi fitur dari POS Tag, unigram dan title case, yang mengungguli 31,57%

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Tahun: 2019 **Penerbit:** ScienceDirect	diperlukan. Oleh karena itu diperlukan ekstraksi entitas untuk memudahkan pencarian informasi dalam hadist.		dari IOB Tag tanpa fitur tambahan.
4.	Named-Entity Recognition on Indonesian Teks beritas using Bidirectional LSTM-CRF **Nama:** Deni Cahya Wintakaa, Moch Arif Bijaksanaa, Ibnu Asror **Tahun:** 2019 **Penerbit:** ScienceDirect	**Objek:** Korpus Teks berita **Masalah:** Sebagian besar NER dilatih untuk menangani teks formal seperti berita, tetapi jika diterapkan pada teks informal seperti pada sosial media twitter kinerjanya buruk.	Penelitian ini mengembangkan model NER dengan menggabungkan metode Bidirectional LSTM dan Conditional Random Field as Architecture + PoS	Model mendapatkan hasil skor F1 terbaik dengan menambahkan kata bertipe FastText, yaitu 86,13% untuk teks berita formal, 81,17% untuk teks berita informal, dan 84,11% untuk teks berita gabungan.
5.	People Entity Recognition in Indonesian Quran Translation with Conditional Random Field Approach **Nama:** Farhan Dzaky Arvianto, Moch Arif	**Objek:** Teks Terjemahan Al quran **Masalah:** Al Quran memiliki total 30 juz, terbagi menjadi 144 surah dan tersusun dari 6.236 ayat dan di	Conditional Random Field	Pada pengembangan sistem untuk mengidentifikasi entitas umat dalam Alquran dengan dataset terjemahan Alquran bahasa Indonesia menunjukkan, pengujian ini menghasilkan rata-rata

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Bijaksana, Arief Fatchul Huda **Tahun:** 2019 **Penerbit:** IEEE	dalam Al Quran membahas topik yang berbeda-beda dan memiliki entitas yang banyak pula, sehingga seseorang terkadang kesulitan memahami Al Quran. Untuk memudahkan dalam memahami Al-Qur'an, kita bisa mendapatkan identifikasi entitas esensial dalam Al-Qur'an seperti nama-nama orang dalam Al-Qur'an dengan menggunakan NER		kinerja menggunakan F1 yang dihasilkan sebesar 0,77 untuk penggunaan data pelatihan sebanyak 36814 data dari 954 ayat Alquran.
6.	Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory **Nama:** Joan Santoso, Esther Irawati Setiawan, Christian Nathaniel	**Objek:** Teks Bahasa Indonesia **Masalah:** Bagaimana mengekstraksi informasi dari kumpulan teks bahasa Indonesia menggunakan NER	Bidirectional long shortterm memory	Dari hasil penelitian didapatkan F1-Score Named Entity Recognition sebesar 83,18%.

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Purwanto, Eko Mulyanto Yuniarno, Mochamad Hariadi, Mauridhi Hery Purnomo **Tahun:** 2021 **Penerbit:** ScienceDirect			
7.	TENER: Adapting *Transformer* Encoder for Named Entity Recognition **Nama:** Hang Yan, Bocao Deng, Xiaonan Li, Xipeng Qiu **Tahun:** 2019 **Penerbit:** ResearchGate	**Objek:** Teks Bahasa Inggris **Masalah:** Performa dari metode sebelumnya yaitu *Bi-LSTM* masih perlu ditingkatkan	Tansformer encoding	Eksperimen dengan 6 dataset NER bahasa Inggris dan empat tugas NER Cina menunjukkan bahwa kinerja TENER mendapatkan tingkat akurasi yang lebih baik dari BiLSTM based
8.	Unified *Transformer* Multi-Task Learning for Intent Classification With Entity Recognition **Nama:** Alberto Benayas, Reyhaneh Hashempour, Damian Rumble, Shoaib Jameel, And Renato Cordeiro De Amorim	**Objek:** Teks Bahasa Inggris **Masalah:** Adanya kelemahan dari penggunaan metode basis RNN baik dalam bentuk LSTM ataupun GRU yaitu proses	Multi-task *Transformer*-based Model	Dengan menggunakan rancangan model yang digunakan diperoleh peningkatan performa NER sebesar 1%

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Tahun: 2021 **Penerbit:** IEEE	pada tahap encoding input yang belum efektif . Pada penelitian ini peneliti mengggabungkan metode komputasi antara (Intent Classification)IC dan NER		
9.	Tuning Multilingual *Transformer*s for Named Entity Recognition on Slavic Languages **Nama:** Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, Alexey Sorokin **Tahun:** 2019 **Penerbit:** ResearchGate, IEEE	**Objek:** Teks Bahasa Slavik **Masalah:** Pengenalan entitas terdahulu hanya dilatih pada dataset tunggal sehingga Perlu memperkaya model dengan melatihnya dengan beberapa tugas secara bersamaan yang membuat representasi kata-katanya lebih banyak fleksibel dan kuat untuk (Intent Classification) IC dan NER	BERT	

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
10.	*Transformer* based named entity recognition for place name extraction from unstructured text **Nama:** Cillian Berragan, Alex Singleton, Alessia Calafiore & Jeremy Morley **Tahun:** 2022 **Penerbit:** International Journal of Geographical Information Science	**Objek:** Teks Bahasa Inggris **Masalah:** Bagaimana mengekstrak informasi geografi dari teks tidak terstruktur	*Transformer*: BERT	Hasil dari performa model NER yang dibuat menunjukkan peningkatan pada ekstraksi entitas untuk nama tempat.
11.	*Transformer*s-based Information Extraction with Limited Data for Domain-Specific Business Documents **Nama :** Minh-Tien Nguyena,b, Dung Tien Lea, Le Thai Linhc **Tahun:** 2021 **Penerbit:** ScienceDirect	**Objek:** Teks dalam dokumen bisnis perusahaan Jepang **Masalah:** Bagaimana mengekstraksi informasi dalam dokumen bisnis perushaan Jepang	*Transformer*: BERT + CNN	Hasil penelitian menunjukkan adanya peningkatan performa ekstraksi sebanyak 15%.
12.	Software requirement-specific entity extraction using *Transformer* models	**Objek:** Teks Bahasa Inggris **Masalah:**	Deep learning-based ML-CRF menggunakan model *Transformer*	Penelitian ini mendapatkan peningkatan akurasi sebesar 4% dan 5% untuk masing-masing

No	Judul Karya Ilmiah, Nama, Tahun Terbit Dan Penerbit	Objek Dan Permasalahan	Metode Penyelesaian	Kinerja
	Nama: Garima Malik, Mucahit Cevik, Swayami Bera, Savas Yildirim, Devang Parikh, Ayse Basar **Tahun:** 2022 **Penerbit:**	Bagaimana mengekstraksi entitas pada dokumen Software Requirements Specifications (SRS) untuk mendapatkan informasi yang diinginkan		dataset DOORS dan SRE
13.	Arabic Named Entity Recognition in Arabic Teks beritas Using BERT-based Models **Nama:** Brahim Ait Benalia, Soukaina Mihia, Nabil Laachfoubia , Addi Ait Mloukb **Tahun:** 2022 **Penerbit:** ScienceDirect	**Objek:** Teks Bahasa Arab dalam twitter **Masalah:** Bagaimana mengekstraksi teks tidak terstruktur pada Bahasa Arab	BERT + *Bi-LSTM* CRF	Berdasarkan 6 kali percobaan yang dilakukan pada penelitian ini didapatka peningkatan performa yang signifikan pada penggunaan *Bi-LSTM* -CRF dan BERT

Berdasarkan dari penjelasan pada bagian latar belakang, penjelasan penelitian terkait, dan uraian pada tabel State of The Art diatas, dapat dilihat bahwa penggunaan beberap metode sebelumya masih memiliki kekurangan pada performa dari akurasi NER dalam bahasa Indonesia yang disebabkan karena metode yang digunakan gagal untuk mengenali entitas yang ada dalam teks. Disamping itu terdapat kelemahan lama yang belum teratasi dengan baik pada NER bahasa

Indonesia yaitu adanya ambiguitas pada kata-kata tertentu yang gagal diidentifikasi pada penelitian sebelumnya karena metode yang dipakai hanya memperhitungkan kata sebelum dan kata setelah yang sedang diidentifikasi tanpa melihat konteks keseluruhan dari teksnya.

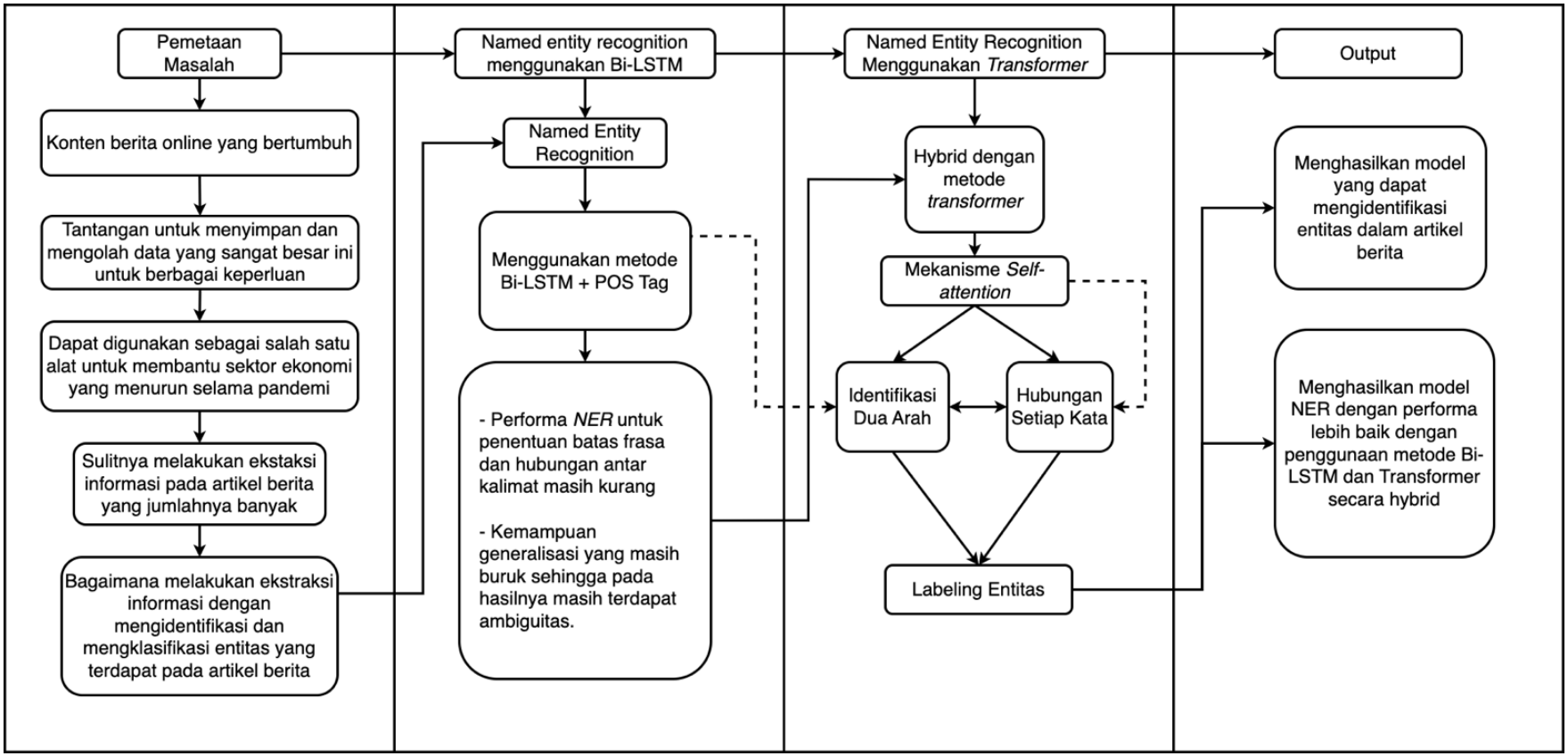
Namun dalam penelitian mengenai ektraksi entitas lain ditemukan hasil bahwa dengan menggunakan metode *Transformer* pada bahasa Inggris dapat mengidentifikasi entitas dengan baik. Hal itu terjadi karena *Transformer* meggunakan self-attention pada proses ektraksinya. Self-attention adalah teknik depending parsing yang digunakan untuk mengidentifikasi hubungan gramatikal antara satu kata dengan kata lain dalam suatu kalimat. Dengan menggunakan dependency parsing, kita dapat memahami struktur gramatikal kalimat dan mengidentifikasi bagaimana setiap kata terhubung dengan kata lain dalam kalimat tersebut. Hal tersebut akhirnya membuat performa akurasi dari NER dapat meningkat dengan baik. Sehingga pada penelitian ini, penulis akan melakukan perancangan dengan ektraksi informasi dengan menggunakan metode kombinasi antara modifikasi metode *Transformer* dan PoS Tagging yang diharapkan mampu mengatasi kelemahan-kelemahan pada penelitian sebelumnya. Dalam penelitian ini penulis juga akan melakukan modifikasi pada model *Transformer* dan PoS Tagging sebagai upaya untuk meningkatkan performa dari kedua metode dan teknik tersebut. Dengan adanya penelitian ini diharapkan dapat mengatasi kekurangan dari penelitian-penelitian sebelumnya pada ekstraksi entitas dalam bahasa Indonesia yang dimana masih terdapat beberapa kekurangan.

2.3 Target Hasil Penelitian

Berdasarkan penelitian yang akan dikerjakan maka target penelitian diharapkan mampu menghasilkan ekstraksi entitas pada berita online Bahasa Indonesia yang memiliki performa lebih baik dari penelitian sebelumnya. Penelitian ini juga diharapkan dapat meningkatkan akurasi dari metode *Transformer* dengan melakukan kostumisasi dengan menerapkan teknik dependencies parsing lain di dalamnya, dalam hal ini Stanford Typed Dependencies. Hasil akhirnya diharapkan performa ekstraksi entitas dalam Bahasa Indonesia dapat meningkat dan mampu

melakukan screening konten pada berita dengan baik dan benar sehingga informasi penting dalam berita bisa didapatkan.

2.4 Kerangka Pikir



Gambar 7 Kerangka Pikir