Jurnal Komputer dan Informatika
Vol 20 No 2, Oktober 2025: hlm 179 – 192

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

# EKSTRAKSI INFORMASI 5W1H DENGAN INDOBERT NAME ENTITY RECOGNITION UNTUK BERITA OLAHRAGA

**George Wielianto**
Program Studi Teknik Informatika, Teknik Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *wieliantog@gmail.com*

## ABSTRAK

Berita olahraga sepak bola, basket, dan bulu tangkis di media online masa kini seringkali mengandung informasi yang sangat kaya di dalam beritanya, seperti nama atlet, tim, kejuaraan, turnamen, hingga informasi-informasi penting seperti skor, stadion sering kali tersembunyi dalam teks tak terstrukur. Penelitian ini bertujuan untuk membantu mengekstraksi informasi menggunakan NER (Named Entity Recognation) dengan model IndoBERT untuk mendapatkan informasi-informasi penting dalam berita olahraga dalam format 5W1H (*What, When, Who, Where, Why*, dan *How*) menggunakan metode rule-based. Model IndoBERT yang telah di fine-tune kemudian akan diuji dengan data uji berita olahraga Indonesia dan memperoleh nilai F1-Score terbaik sebesar 0.8191 pada epoch ke-10, menunjukkan bahwa kinerja dan performa dari model IndoBERT ini bekerja dengan baik dalam mengenali entitas-entitas NER.

**Kata kunci:** *Named Entity Recognition, 5W1H, IndoBERT, Olahraga, rule-base.*

## *ABSTRACT*

*Football, basketball, and badminton sports news in today's online media often contain very rich information in the news, such as the names of athletes, teams, championships, tournaments, to important information such as scores, stadiums are often hidden in unstructured text. This study aims to help extract information using NER (Named Entity Recognition) with the IndoBERT model to obtain important information in sports news in the 5W1H format (What, When, Who, Where, Why, and How) using a rule-based method. The fine-tuned IndoBERT model will then be tested with Indonesian sports news test data and obtained the best F1-Score of 0.8191 in the 10th epoch, indicating that the performance and performance of this IndoBERT model works well in recognizing NER entities*.

***Keywords:*** *Named Entity Recognition, 5W1H, IndoBERT, Sports, rule-based.*

## 1. PENDAHULUAN

Internet telah menjadi bagian penting dari penyebaran informasi dan telah menjadi bagian penting dari kehidupan sebagian besar orang di seluruh dunia. Sebuah laporan yang dibuat oleh Asosiasi Penyelenggara Jasa Internet (APJII) pada tahun 2024 memperkirakan bahwa jumlah orang yang menggunakan internet akan mencapai 221 juta orang, atau sekitar 79,5% dari populasi Indonesia, pada tahun tersebut [1]. Angka ini menunjukkan bahwa pertumbuhan pesat dalam penggunaan internet akan membantu pemerataan informasi di masyarakat [2]. Media online telah berkembang menjadi sumber informasi penting, termasuk berita ekonomi, politik, dan olahraga. Lebih dari 50% waktu masyarakat Indonesia dihabiskan di situs berita dan platform streaming, menurut penelitian yang dilakukan oleh media Indonesia. Karena generasi ini tumbuh di era digital yang menyebabkan semakin banyak orang memilih mengakses berita melalui media daring dengan alasan kemudahan akses dan informasi yang cepat [3].

Meskipun informasi semakin mudah diakses, minat baca masyarakat Indonesia masih rendah. Sebuah penelitian yang dilakukan oleh UNESCO menunjukkan bahwa hanya 0,001% populasi Indonesia menyukai membaca [4]. Menurut penelitian yang dilakukan oleh News and Media Research Centre (N&MRC) University of Canberra, orang Indonesia lebih suka membaca ringkasan berita yang dibuat oleh AI daripada membaca artikel secara keseluruhan [5].

Sistem ekstraksi informasi yang memanfaatkan teks berita untuk menyampaikan informasi penting tanpa harus membaca artikel secara keseluruhan didasarkan pada masalah minat baca yang rendah. Informasi penting seperti nama atlet, tim, stadion, kejuaraan, dan skor pertandingan seringkali tersembunyi dalam teks berita olahraga yang tidak terstruktur. Pendekatan Named Entity Recognition (NER) mengidentifikasi entitas seperti nama orang, waktu, tempat, dan lain-lain, dan digunakan untuk mengekstrak informasi ini [6]. Model Bidirectional Encoder Representations from Transformers (BERT), model pra-latihan dengan arsitektur Transformer yang dikembangkan oleh Google [7], akan digunakan dalam perancangan dan pengembangan sistem ekstraksi ini. Model IndoBERT, sebuah adaptasi dari BERT yang telah dilatih menggunakan korpus bahasa Indonesia yang besar [8].

Penelitian sebelumnya telah dilakukan untuk menciptakan sistem NER yang dapat mengekstrak data dalam berbagai domain dan bahasa. Sebagai contoh, Chantrapornchai dan Tunasakul (2019) menggunakan model BERT untuk mengidentifikasi entitas seperti nama, lokasi, dan fasilitas dari teks Inggris tentang pariwisata [9]. Selain itu, Darji dan Mitrovic (2023) menggunakan model BERT untuk menemukan entitas hukum dalam teks Jerman, menunjukkan bahwa model dapat disesuaikan dengan konteks bahasa dan domain tertentu [10]. Penelitian menunjukkan bahwa perancangan ini berfokus pada pembuatan sistem ekstraksi informasi untuk berita olahraga Indonesia. Ini akan menggunakan model IndoBERT dan diintegrasikan dengan rule-based untuk menampilkan format yang lebih terstruktur.

## 2. TINJAUAN LITERATUR

Berdasarkan penelitian mengenai Named Entity Recognation yang telah dilakukan pada berbagai domain dan bahasa dengan maksud tujuan untuk membantu dan menjadi dasar dalam perancangan ini.

Studi tentang NER menunjukkan bahwa model berbasis transformer, khususnya BERT dan variasinya, meunjukkan peningkatan di dalam banyak hal. Studi Tunsakul (2020) pada korpus pariwisata menunjukkan bahwa model BERT mengungguli metode yang lebih konvensional seperti CRF dalam memahami konteks kata, yang memungkinkan pengenalan entitas yang lebih akurat untuk lokasi, amenitas, dan atraksi [9]. BERT, yang dikembangkan oleh Darji dan Mitrovic (2023), menunjukkan peningkatan kinerja berbasis domain. Studi mereka menunjukkan bahwa BERT, yang dilatih secara khusus pada domain hukum, lebih baik daripada model multibahasa dalam penelitian mereka [10]. Ini menunjukkan betapa pentingnya model yang dioptimalkan untuk domain dan bahasa.

Koto dkk. (2020) menciptakan model pra-latihan dan standar untuk berbagai tugas NLP di Indonesia dengan IndoLEM dan IndoBERT [11]. Karena tidak ada penekanan khusus pada domain tertentu, beberapa entitas kurang terwakili meskipun cakupannya luas. Penelitian oleh Wilie dkk. (2020) menggunakan IndoNLU menunjukkan bahwa IndoBERT sering mengungguli mBERT dalam beberapa tugas NLP, termasuk tugas NER, meskipun dataset NER Indonesia sangat terbatas dan belum mencakup semua domain khusus, seperti olahraga atau hukum.

Menurut penelitian yang dilakukan oleh Liu et al. (2020) dalam bidang berita Esports, teknik gabungan CRF dan BERT dapat mengekstrak entitas dari teks pertandingan Esports. Meskipun demikian, kinerjanya, yang hanya mencapai sekitar 61% dari skor F1, menunjukkan bahwa sistem berbasis NER masih menghadapi masalah dalam domain tertentu dengan struktur bahasa yang sangat beragam [12]. Jing Li et al. (2020) memberikan analisis menyeluruh tentang kemajuan teknik pembelajaran mendalam untuk NER [6]. Meskipun studi ini merupakan survei dan tidak melakukan uji coba pada bahasa Indonesia, ia menunjukkan bahwa pendekatan berbasis Transformer adalah standar terbaru untuk ekstraksi entitas.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Menurut enam penelitian, model pra-latih berbasis transformer seperti BERT dan IndoBERT dapat memahami konteks dan mengekstrak entitas dari berbagai domain dengan baik. Namun demikian, sebagian besar penelitian hanya membahas klasifikasi entitas dan gagal mengintegrasikan hasil NER ke dalam struktur informasi yang canggih seperti 5W1H. Studi sebelumnya belum menggunakan pendekatan berbasis aturan untuk mengintegrasikan hasil NER ke dalam kategori Siapa, Apa, Kapan, Di Mana, Mengapa, dan Bagaimana. Akibatnya, integrasi sistem berbasis aturan dengan IndoBERT, yang telah disempurnakan dengan berita olahraga Indonesia, menghasilkan ekstraksi informasi yang lebih terorganisir dan mudah dipahami.

Selain itu, penelitian sebelumnya sering melihat kinerja NER sebagai tugas tunggal tanpa mengembangkannya ke interpretasi semantik yang lebih baik. Penelitian tertentu menggunakan metode pasca-pemrosesan, tetapi metode ini terbatas pada normalisasi atau penautan entitas daripada penalaran hierarkis, yang mencakup pemetaan entitas ke Siapa, Apa, Kapan, Di Mana, Mengapa, dan Bagaimana. Akibatnya, studi sebelumnya belum membahas kebutuhan untuk mengonversi keluaran entitas mentah menjadi format yang lebih ramah pengguna dan bermakna secara kognitif. Perbedaan ini sangat penting dalam industri seperti pemrosesan berita, di mana pengguna seringkali membutuhkan ringkasan terperinci daripada teks lengkap. Oleh karena itu, belum ada penelitian sebelumnya yang secara metodis mengintegrasikan lapisan penalaran berbasis aturan yang dirancang khusus untuk menerjemahkan label entitas ke dalam struktur 5W1H dengan model NER yang telah disempurnakan. Ini menunjukkan inovasi dan kontribusi studi ini, yang menjembatani kesenjangan ini dengan menggabungkan sistem pemetaan berbasis aturan deterministik dengan NER berbasis IndoBERT untuk menghasilkan hasil ekstraksi informasi yang terstruktur dan mudah dipahami.

## 3. METODE PENELITIAN

Pada bagian ini akan berfokus pada metode penelitian yang akan dipakai, yang mencakup BERT sebagai pendekatan umum dan model yang akan digunakan dalam rancangan ini, yang bertujuan untuk mengembangkan dan menguji sistem ekstraksi informasi pada berita olahraga Indonesia menggunakan metode *Named Entity Recognition* (NER). Pada perancangan ini dilakukan dengan beberapa tahapan yaitu mulai dari desain sistem, persiapan dataset, serta pembuatan sistem hingga implementasi ke dalam aplikasi website untuk menampilkan hasil ekstraksi dalam format 5W1H (*Who, What, Where, When, Why*, dan *How*).

### 3.1 Desain Sistem

Pada tahap desain sistem ini bertujuan untuk memberikan gambaran secara umum mengenai rancangan sistem ekstraksi informasi berita olahraga yang akan dikembangkan. Proses perancangan ini akan melibatkan empat komponen utama pada perancangan ini yaitu, *flowchart* sistem, skema algoritma *rule-based*, desain website, dan rancangan dataset berita olahraga yang akan digunakan sebagai data latih model IndoBERT.

Proses perancangan sistem ekstraksi ini diawali dengan pengumpulan dataset dari tiga media online berita yang ada di Indoneisa yaitu DetikSport, KompasSport, dan BolaSport. Lalu dataset yang sudah dikumpulkan akan dilabeli atau diberikan anotasi entitas-entitas NER pada domain olahraga secara manual menggunakan software *open-source* yaitu LabelStudio. Setelah semua dataset berita yang akan dipakai dilabeli kemudian dataset tersebut dibagi menjadi tiga bagian yaitu menjadi dataset latih, validasi, dan uji yang akan digunakan untuk proses pelatihan *fine-tuning* model IndoBERT agar dapat melakukan deteksi NER pada entitas-entitas olahraga. Selanjutnya proses pelatihan model IndoBERT akan dilakukan menggunakan *hyperparameter* yang sudah ditetapkan dengan menggunakan *loss function* BCEWithLogitsLoss dan optimizer AdamW. Setelah model selesai dilatih, model akan melakukan evaluasi dengan data uji ynag sudah dibagi sebelumnya untuk mengukur performa dan akurasi hasil *fine-tuning* dari model IndoBERT.

Lalu setelah model IndoBERT sudah dilatih, pada perancangan ini juga akan menggunakan algoritma *rule-based* untuk melakukan pemetaan terhadap label-label entitas ke dalam format 5W1H yang terstruktur dan mudah dibaca. Algoritma *rule-based* ini bekerja dengan cara menggabungkan token-token yang berurutan dan memikili label yang sama agar membentuk satu entitas utuh, misalnya "Lionel" (ATLET) dan "Messi" (ATLET) digabung menjadi "Lionel Messi". Setelah itu setiap entitas akan di*mapping* ke dalam kategori 5W1H (What, Who, When, Where, Why, dan How). Pemetaan kedalam format 5W1H dapat dilihat pada Tabel 1.

**Tabel 1** Pemetaan Entitas ke Dalam 5W1H

| No | Label NER | Kategori 5W1H |
|---|---|---|
| 1. | ATLET | WHO |
| 2. | TIM | WHO |
| 3. | ORGANISASI | WHO |
| 4. | KEJUARAAN | WHERE |
| 5. | STADION | WHERE |
| 6. | STATISTIK | WHAT |
| 7. | PENGHARGAAN | WHAT |
| 8. | POSISI | WHO |
| 9. | KEWARGANEGARAAN | WHO |
| 10. | UMUR | WHO |
| 11. | TANGGAL | WHEN |
| 12. | SKOR | WHAT |
| 13. | AKSI | HOW |
| 14. | ALASAN_PERISTIWA | WHY |

Lalu pada tahap terakhir desain sistem ini dilakukan desain sistem aplikasi website sistem ekstraksi informasi yang dirancangan dengan *user interface* yang sederhana, responsif, dan *user friendly* agar dapat dengan mudah digunakan oleh semua kalangan. Website yang akan dibangun bernama SportExtract terdiri dari empat komponen utama yaitu: (1) *header*, bagian ini meliputi nama aplikasi kemudian berisikan informasi tentang media online apa saja yang didukung dan kategori berita yang didukung, (2) kolom input berita disisi kiri aplikasi website dengan sistem *tab switching* untuk membedakan metode input teks berita dan URL berita, (3) kemudian terdapat juga contoh link pada kolom input URL berita, contoh link berita diambil dari media online yang didukung yaitu DetikSport, KompasSport, dan BolaSport, (4) kolom output di sisi kanan website yang menampilkan hasil ekstraksi informasi dari berita yang sudah diproses dalam format 5W1H. alur kerja aplikasi website ini akan dimulai dengan input teks berupa teks berita atau URL berita, kemudia jika input berupa URL berita aplikasi ini akan melakukan scrapping otomatis untuk mendapatkan isi konten berita tersebut dan kemudian akan melakukan *preprocessing*. Selanjutnya teks berita yang sudah bersih akan diproses menggunakan model IndoBERT yang sudah di *fine-tune* sebelumnnya untuk melakukan deteksi NER, setelah model melakukan deteksi NER hasil dair entitas-entitas tersebut kemudian akan dipetakan ke dalam 5W1H yang akan dijadikan sebagai output. Untuk melihat alur proses dari sistem aplikasi website ini dapat dilihat pada Gambar 1.


**Gambar 1** Flowchart Sistem (Sumber: Dokumentasi Pribadi)

3.2 Persiapan Dataset

Sebelum membuat sistem ekstraksi ini dilakukan persiapan dataset untuk dijadikan dataset pada proses pelatihan, validasi, dan pengujian model IndoBERT. Dataset yang akan digunakan pada rancangan sistem ini adalah dataset yang berisikan kumpulan berita-berita olahraga Indonesia yang diambil dari tiga cabang olahraga yaitu sepakbola, basket, dan badminton.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Proses pengumpulan dataset ini dilakukan dengan cara menggunakan *framework* Scrapy yang menggunakan *library* BeautifulSoup untuk melakukan ekstraksi konten-konten teks berita dari halaman website media online dalam rentang waktu 1 tahun (1 Januari 2024 sampai 30 Agustus 2025) Setiap kode scrapping akan dibuat secara terpisah untuk masing-masin situs berita agar dapat menyesuaikan struktur HTML yang berbeda-beda. Proses scrapping ini dilakukan dengan cara:

1. Mengambil daftar URL artikel dari halaman utama dan subhalaman kategori olahraga.
2. Menelusuri setiap link artikel dan melakukan ekstraksi informasi seperti judul, tanggal, publikasi, isi berita, dan link dari sumber media.
3. Membersihkan elemen-elemen yang tidak relevan seperti iklan, rekomendasi artikel, dan link-link yang tidak relevan lainnya, sehingga teks konten berita yang disimpan dalam file .csv adalah dataset yang bersih.
4. Menyimpan hasil ekstrasi ke dalam format file .json.

Dataset yang sudah terkumpul dalam file .json lalu akan diimport ke dalam LabelStudio dan akan diberi anotasi terlebih dahulu menggunakan sistem *software open-source* ini. Setiap gambar nantinya akan diberikan label entitas NER pada domain olahraga. Proses pelabelan data ini juga akan dilakukan secara manual untuk menjaga kekonsistenan dari setiap label. Setelah melakukan anotasi menggunakan LabelStudio dataset berita yang sudah dilabeli akan di export ke dalam format span-based .csv dan siap untuk dijadikan dataset utama untuk melatih model NER. Format *span-based* ini akan menyimpan informasi dari indeks awal dan indeks akhir dari setiap entitas. Format ini dipakai untuk menangani masalah *overlapping entities*. Verifikasi oleh pakar dibidang junalistik juga dilakukan untuk memperkuat validasi data dan hasil dari pelabelan manual ini.

Setelah proses pengumpulan dan pelabelan data berita selesai dilakukan kemudian saat tahap pelatihan nantinya dataset akan dibagi menjadi tiga bagian yaitu data latih (80%), data validasi (10%), dan data uji (10%). Jumlah dataset yang dikumpulkan bisa dilihat pada Tabel 2.

**Tabel 2** Detail Jumlah Dataset

| Media | Cabor | Jumlah Data |
|---|---|---|
| DetikSport | Basket | 233 |
| | Sepakbola | 234 |
| | Bulutangkis | 233 |
| KompasSport.com | Basket | 233 |
| | Sepakbola | 234 |
| | Bulutangkis | 233 |
| BolaSport.com | Basket | 233 |
| | Sepakbola | 234 |
| | Bulutangkis | 233 |
| Total | | 2100 |

### 3.3 Pembuatan Sistem

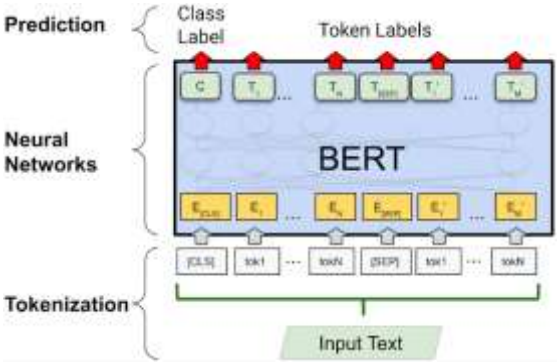
Selanjutnya proses pembuatan sistem ini akan dilakukan menggunakan bahasa pemrograman python dan framework flask yang akan digunakan sebagai backend. Model IndoBERT yang telah dilatih dan diuji lalu akan diintegrasikan dengan aplikasi website. Sistem ini juga sudah mengintegrasikan scrapping otomatis menggunakan library BeautifulSoup untuk mengekstrak isi berita dari input URL yang dimasukkan oleh pengguna.

#### 3.3.1 Pelatihan Model NER

*Bidirectional Encoder from Transformers* (BERT) adalah salah satu model pre-trained language model yang diperkenalkan oleh Devlin et al. model BERT ini menggunakan arsitektur dasar Transformer, terutama pada bagian encodernya, yang berfungsi untuk menghasilkan representasi teks yang kaya akan konteks secara dua arah (*bidirectional*). Berbeda dengan model-model sebelumnya seperti *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM),

Transformer sendiri tidak bergantung pada pemrosesan data secara berurutan [11]. Arsitektur ini menggunakan mekanisme self-attention untuk memahami hubungan antar kata dalam sebuah kalimat secara paralel dan lebih efisien dalam memahami konteks yang panjang [12].

Arsitektur Transformer terdiri dari dua komponen utama yaitu encoder dan decoder. Namun pada model LLM BERT ini hanya digunakan komponen encoder karena fokus utama dari model ini sendiri adalah pemahaman teks. Setiap layer encoder ini memiliki komponen-komponen inti seperti *Token Embedding* dan *Positional Encoding, Multi-Head Self-Attention*, dan *Feedforward Neural Network* [13].




**Gambar 2** Arsitektur Transformer [11]

Model yang akan digunakan pada perancangan dan pembuatan sistem ini adalah IndoBERT base-p1 dari IndoBERT Benchmark, yaitu model BERT yang telah dilatih khusus untuk bahasa Indonesia. IndoBERT-base ini adalah model dasar dari IndoBERT, yang telah dilatih dengan korpus 5,5 miliar kata yang mencakup beberapa bentuk teks bahasa Indonesia. Model ini dapat digunakan untuk tugas NLP. Model ini sendiri terdiri dari 12 lapisan *transformator* dengan 12 *heads* per lapisan dan 110 juta parameter [14].

Model di *fine-tune* menggunakkan dataset berita olahraga yang sudah dilabeli dengan entitas-entitas pada domain olahraga. pelatihan ini akan dilakukan menggunakan *PyTorch* dan *Library Transformers* dengan hyperparameter yang ditetapkan sebagai berikut:

**Tabel 3** Hyperparamter yang Dipakai

| Hyperparameter | Nilai |
|---|---|
| Model dasar | indobenchmark/indobert-base-p1 |
| Batch size | 2 |
| Learning rate | 2e-5 |
| Weight decay | 0.01 |
| Epoch | 15 |
| Max_length | 512 |
| Optimizer | AdamW |

Selama proses pelatihan ini model akan menghasilkan *checkpoint* pada setiap epoch untuk memilih model terbaik berdasarkan nilai *F1-Score* yang tertinggi pada proses pelatihan model.

### 3.3.2 Rule-Based 5W1H

Pada perancangan ini setelah model IndoBERT menghasilkan entitas-entitas NER lalu hasil dari entitas-entitas tersebut akan dilakukan pemetaan ke dalam format 5W1H dengan menggunakan metode *rule-based*. Pemetaan ini dilakukan dengan cara menggabungkan 5W1H dengan tipe-tipe entitas, pemetaan entitas dapat dilihat pada **Error! Reference source not found.**Dengan pendekatan *rule-based* ini berguna untuk menghasilkan hasil ekstraksi yang lebih rapih, terstruktur dan mudah dibaca.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

3.4 Pengujian Sistem

Tahapan ini akan dilakukan untuk memverifikasi dan memastikan bahwa sistem ekstraksi berjalan dengan baik dan sesuai fungsinya. Pengujian akan mencakup tiga aspek yaitu:

a. Pengujian performa IndoBERT menggunakan evaluasi metrik *Precision, Recall,* dan *F1-Score.*
b. Pengujian menggunakan *human evaluation*, dimana sistem ekstraksi ini akan diuji oleh masyarakat.

3.4.1 Evaluasi Metrik

Pada rancangan ekstraksi informasi pada berita olahraga Indonesia ini terdapat 14 entitas yang akan ekstraksi, maka dari itu digunakan evaluasi metrik untuk melakukan penilaian terhadap hasil ekstraksi dari model yang sudah dilatih. Metrik yang digunakan yaitu *Precision, Recall,* dan *F1-Score*, Untuk mengukur performa dari model yang akan digunakan.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (1)$$

*Precision* digunakan untuk mengukur tingkat keakuratan dan ketepatan hasil prediksi entitas yang dilakukan oleh model IndoBERT. *Precision* ini akan menilai seberapa banyak entitas yang diprediksi dengan benar dibandingkan dengan entitas yang diprediksi sebagai positif. Sebagai contoh jika model memprediksi 100 entitas "ATLET" dan 80 diatanranya benar dengan label yang sebenarnya, maka nilai *precision* sebesar 0,8 [15].

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

*Recall* mengukur sebarapa jauh model berhasil menemukan seluruh entitas yang seharusnya ditemukan. Metrik ini akan menghitung jumlah entitas yang berhasil diprediksi dengan benar terhadap total entitas yang sebnarnya ada dalam data. Contohnya jika dalam berita terdapat 100 entitas "ATLET" dan prediksi model hanya berhasil memprediksi 85. Maka nilai recall adalah 0,85 [15].

$$F-1\,Score = 2 * \frac{\mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \quad (3)$$

*F1-Score* merupakan rata-rata antara *precision* dan *recall* yang akan dibobotkan. Metrik ini berfungsi untuk memberikan nilai keseimbangan antara kedua metrik *precision* dan *recall* [15].

3.4.2 Evaluasi Manusia

Selain evaluasi metrik, pada perancangan ini juga dilakukan evaluasi manual terhadap hasil deteksi entitas yang akan dihasilkan oleh model. Dalam tahap ini akan diambil sejumlah sampel teks berita yang akan diuji dengan hasil output dari model dan dibandingkan dengan anotasi asli oleh responden.

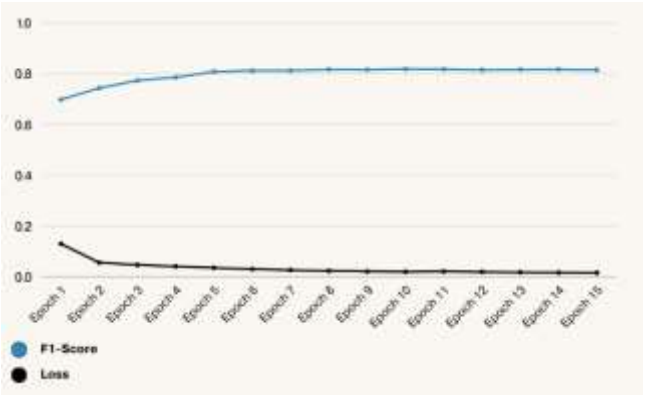
## 4. HASIL DAN PEMBAHASAN

Selanjutnya pada bagian ini, menunjukkan hasil dari kinerja model yang sudah dirancangan dan dibuat denga *hyperparameter* yang sudah ditentukan. Hasil yang ditunjukkan ini akan menggunakan model *pretrained* IndoBERT-base-p1 yang telah di *fine-tune* sebelumnya, yang diperoleh dari repositori Indo *Benchmark* pada *Hugging Face*. Cara pengujian sistem ekstraksi informasi ini menggunakan pengujian evaluasi metrik untuk menguji model IndoBERT yang digunakan.

4.1 Hasil Pengujian Model IndoBERT

Pengujian model ini dilakukan dengan cara menggunakan 100 data uji berita olahraga Indonesia yang belum pernah dilihat pada data latih sebelumnnya. Berdasarkan hasil dari pengujian

ini nilai *F1-Score* paling baik diperoleh dengan nilai sebesar 0.8191 pada epoch ke-10. Sedangkan pada epoch-epoch berikutnya nilai *F1-Score* terus menurun, hal ini terjadi karena setelah epoch ke-10 model mulai mengalami *overfitting*. Grafik nilai *F1-Score* dan loss dapat dilihat pada gambar 3. Evaluasi ini dilakukan kepada seluruh 14 entitas NER. Sehingga hasil dari *F1-Score* dan *loss* ini akan menghasilkan performa keseluruhan dari model ini sendiri dalam mengenali semua kategori entitas pada berita olahraga Indonesia.




**Gambar 3** *Grafik F1-Score* dan *Loss*

Selain melakukan evaluasi F1-Score terhadap seluruh label, pada perancangan ini juga dilakukan evaluasi terhadap masing-masing label atau entitas menggunakan metrik *precision, recall,* dan *F1-Score*. Evaluasi ini bertujuan untuk menilai seberapa jauh model IndoBERT untuk mengenali setiap label-label entitas secara masing-masing. Hasil dari evaluasi masing-masing label dapat dilihat pada

(a) Label TIM

Pada Label TIM ini menunjukkan performa model dalam medeteksi entitas TIM. Dapat dilihat nilai F1-Score meningkat secara bertahap hingga mencapai stabilitas di tengah epoch (sekitar epoch ke-8) kemudian sedikit menurun di akhir epoch. Hal ini menunjukkan bahwa model IndoBERT cukup baik untuk mengenali entitas TIM pada berita olahraga, namun masih terdapat kesalahan pada hasil deteksinya juga.

(b) Label ATLET

Pada Label ATLET ini menunjukkan peningkatan yang konsisten dan stabil dengan nilai F1-Score rata-rata 0.81-0.82. hal ini menunjukkan bahwa model sangat baik dalam mengenali entitas ATLET, karena nama atlet sering muncul berulang dalam dataset sehingga model dapat belajar dengan maksimal dan baik.

(c) Label KEJUARAAN

Pada Label KEJUARAAN ini menunjukkan peningkatan yang signifikan di awal hingga pertengahan epoch. Model dapat belajar untuk mengenali entitas KEJUARAAN pada berita olahraga Indonesia di bidang Sepakbola, Basket, dan Badminton.

(d) Label STADION

Pada Label STADION ini menunjukkan peningkatan secara bertahap dengan nilai F1-Score rata-rata 0.70-0.80. menurut tabel ini model mampu mengenali entitas STADION dengan baik dan stabil.

(e) Label ALASAN_PERISTIWA

Pada laebl ALASAN_PERISTIWA, nilai F1-Score untuk entitas ALASAN_PERISTIWA adalah yang paling tinggi, karena model berhasil mengenali struktur alasan peristiwa dalam berita olahraga yang selalu diambil dari kalimat pertama yang selalu dijadikan tagline oleh media-media

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

online. Karena polanya yang konsisten maka model dapat belajar dan menhasilkan evaluasi yang cukup baik.

(f) Label AKSI

Pada label AKSI ini dapat dilihat hasil evaluasi lebih stabil dipertengahan epoch. Entitas AKSI ini memiliki variasi frasa yang sangat luas biasanya 1 kalimat penuh sehingga walau hasil evaluasi cukup baik masih ada potensi model masih bisa salah untuk menebak.

(g) Label POSISI

Pada label POSISI dapat dilihat bahwa nilai evaluasi yang didapatkan pada awal epoch sangat rendah yang kemudian dapat meningkat perlahan-lahan hingga epoch terkahir. Nilai F1-Score dari posisi ini juga disebabkan oleh konteks posisi seperti “Striker”, “shooting guard”, dan “Bek tengah” yang sangat bervariasi.

(h) Label UMUR

Pada label UMUR nilai F1-Score cenderung stabil mulai dari epoch ke-3, model mulai mengenali pola numerik dengan kata “tahun” seperti contoh “Bek Spanyol 29 tahun”. Sehingga entitas umur dapat nilai evaluasi yang stabil.

(i) Label ORGANISASI

Pada label ORGANISASI, pada tabel menunjukkan peningkatan nilai F1-Score mulai epoch pertama sampai terakhir. Label organisasi seperti “FIFA”, “NBA” dapat dikenali dengan baik, namun masih belum sempurna.

(j) Label KEWARGANEGARAAN

Pada label KEWARGANEGARAAN, performa model pada label KEWARGANEGARAAN ini dilihat sangat stabil dengan rata-rata F1-Score 0.5. hal ini juga disebabkan karena entitas kewarganegaraan seperti "pemain Spanyol”, “pebulutangkis asal Indonesia” juga jarang muncul pada berita.

(k) Label TANGGAL

Pada Label TANGGAL, performa model pada label TANGAL mengalami peningkatan secara signifikan sejak epoch awal dan mencapai nilai F1-Score tertinggi yaitu 0.9162. Hal ini dapat disebabkan karena tanggal memiliki pola yang konsisten sehingga model mudah untuk memprediksi dan mengenalinya.

(l) Label SKOR

Pada Label SKOR nilai F1-Score pada entitas skor adalah yang paling tinggi, ini dikarenakan model sangat mengenali entitas SKOR karena pola angka seperti 2-0, 3-0 sangat terstruktur dan sering muncul pada dataset.

(m) Label PENGHARGAAN

Pada label PENGHARGAAN ini hanya mendapatkan F1-Score tertinggi 0.6776 pada epoch ke-5, hal ini disebabkan karena entitas penghargaan seperti “MVP”, “Pemain Terbaik” jarang muncul pada dataset berita olahraga.

(n) Label STATISTIK

(a-n) yang menunjukkan hasil performa dari model IndoBERT. Dari hasil evaluasi tersebut, dapat diidentifikasi juga label entitas mana yang memiliki nilai yang paling tinggi dan label entitas mana yang masih memerlukan peningkatan.


TIM

(a)


ATLET

(b)


KEJUARAAN

(c)


STADION

(d)


ALASAN_PERISTIWA

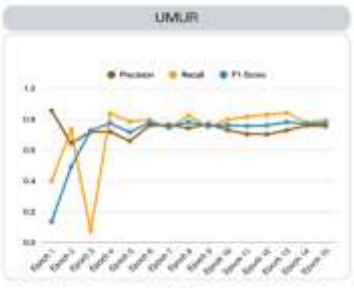
(e)


AKSI

(f)


POSISI

(g)


UMUR

(h)


ORGANISASI

(i)


KEWARGANEGARAAN

(j)

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)


**Gambar 4.** Hasil Evaluasi Masing-masing Label

(a) Label TIM

Pada Label TIM ini menunjukkan performa model dalam medeteksi entitas TIM. Dapat dilihat nilai F1-Score meningkat secara bertahap hingga mencapai stabilitas di tengah epoch (sekitar epoch ke-8) kemudian sedikit menurun di akhir epoch. Hal ini menunjukkan bahwa model IndoBERT cukup baik untuk mengenali entitas TIM pada berita olahraga, namun masih terdapat kesalahan pada hasil deteksinya juga.

(b) Label ATLET

Pada Label ATLET ini menunjukkan peningkatan yang konsisten dan stabil dengan nilai F1-Score rata-rata 0.81-0.82. hal ini menunjukkan bahwa model sangat baik dalam mengenali entitas ATLET, karena nama atlet sering muncul berulang dalam dataset sehingga model dapat belajar dengan maksimal dan baik.

(c) Label KEJUARAAN

Pada Label KEJUARAAN ini menunjukkan peningkatan yang signifikan di awal hingga pertengahan epoch. Model dapat belajar untuk mengenali entitas KEJUARAAN pada berita olahraga Indonesia di bidang Sepakbola, Basket, dan Badminton.

(d) Label STADION

Pada Label STADION ini menunjukkan peningkatan secara bertahap dengan nilai F1-Score rata-rata 0.70-0.80. menurut tabel ini model mampu mengenali entitas STADION dengan baik dan stabil.

(e) Label ALASAN_PERISTIWA

Pada laebl ALASAN_PERISTIWA, nilai F1-Score untuk entitas ALASAN_PERISTIWA adalah yang paling tinggi, karena model berhasil mengenali struktur alasan peristiwa dalam berita olahraga yang selalu diambil dari kalimat pertama yang selalu dijadikan tagline oleh media-media online. Karena polanya yang konsisten maka model dapat belajar dan menhasilkan evaluasi yang cukup baik.

(f) Label AKSI

Pada label AKSI ini dapat dilihat hasil evaluasi lebih stabil dipertengahan epoch. Entitas AKSI ini memiliki variasi frasa yang sangat luas biasanya 1 kalimat penuh sehingga walau hasil evaluasi cukup baik masih ada potensi model masih bisa salah untuk menebak.

(g) Label POSISI

Pada label POSISI dapat dilihat bahwa nilai evaluasi yang didapatkan pada awal epoch sangat rendah yang kemudian dapat meningkat perlahan-lahan hingga epoch terkahir. Nilai F1-Score dari posisi ini juga disebabkan oleh konteks posisi seperti “Striker”, “shooting guard”, dan “Bek tengah” yang sangat bervariasi.

(h) Label UMUR

Pada label UMUR nilai F1-Score cenderung stabil mulai dari epoch ke-3, model mulai mengenali pola numerik dengan kata “tahun” seperti contoh “Bek Spanyol 29 tahun”. Sehingga entitas umur dapat nilai evaluasi yang stabil.

(i) Label ORGANISASI

Pada label ORGANISASI, pada tabel menunjukkan peningkatan nilai F1-Score mulai epoch pertama sampai terakhir. Label organisasi seperti “FIFA”, “NBA” dapat dikenali dengan baik, namun masih belum sempurna.

(j) Label KEWARGANEGARAAN

Pada label KEWARGANEGARAAN, performa model pada label KEWARGANEGARAAN ini dilihat sangat stabil dengan rata-rata F1-Score 0.5. hal ini juga disebabkan karena entitas kewarganegaraan seperti "pemain Spanyol”, “pebulutangkis asal Indonesia” juga jarang muncul pada berita.

(k) Label TANGGAL

Pada Label TANGGAL, performa model pada label TANGAL mengalami peningkatan secara signifikan sejak epoch awal dan mencapai nilai F1-Score tertinggi yaitu 0.9162. Hal ini dapat disebabkan karena tanggal memiliki pola yang konsisten sehingga model mudah untuk memprediksi dan mengenalinya.

(l) Label SKOR

Pada Label SKOR nilai F1-Score pada entitas skor adalah yang paling tinggi, ini dikarenakan model sangat mengenali entitas SKOR karena pola angka seperti 2-0, 3-0 sangat terstruktur dan sering muncul pada dataset.

(m) Label PENGHARGAAN

Pada label PENGHARGAAN ini hanya mendapatkan F1-Score tertinggi 0.6776 pada epoch ke-5, hal ini disebabkan karena entitas penghargaan seperti “MVP”, “Pemain Terbaik” jarang muncul pada dataset berita olahraga.

(n) Label STATISTIK

Pada label STATISTIK mendapatkan nilai evaluasi sangat rendah (0.0) di awal epoch namun meningkat perlahan-lahan hingga stabil sampai akhir epoch. Hal ini terjadi karena STATISTIK ini cukup kompleks karena mencakup kalimat-kalimat yang mengandung “pada menit ke-“, “14 gol, 13 assist” dan lainnya, sehingga model memerlukan data yang lebih banyak lagi untuk mendapatkan hasil yang lebih baik.

Secara keseluruhan dari hasil evaluasi yang dilakukan dengan masing-masing label dapat dilihat bahwa setiap label entitas memiliki nilai *precision, recall* dan *F1-Score* yang bervariasi, hal ini dapat terjadi karena disebabkan oleh tingkat kompleksitas dan jumlah data pelatihan yang mewakili setiap label entitas. Secara umum model IndoBERT sudah menunjukkan performa yang cukup baik dalam mengenali setiap entitas dengan nilai *F1-Score* di atas 0.8.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Dari hasil keseluruhan pada perancangan ini disimpulkan bahwa model IndoBERT mampu melakukan ekstraksi dengan baik pada label-label yang sudah ditentukan, seperti nama atlet, tim, skor dan lainnya. Berdasarkan hasil evaluasi dapat disimpulkan bahwa model sudah mampu untuk melakukan ekstraksi informasi dengan pendekatan NER.

## 5. KESIMPULAN

Pada penelitian ini menunjukkan bahwa model IndoBERT yang telah mengalami fine-tuning menggunakan dataset Indonesia dapat melakukan deteksi entitas (NER) dengan kinerja yang sangat baik, terbukti dari tingginya F1-Score sebesar 0,8191 pada epoch ke-10. Hal ini menunjukkan bahwa model tersebut dapat memahami pola bahasa dalam teks olahraga dan membahas 14 jenis entitas berbeda yang telah diidentifikasi. Kinerja puncak terlihat pada epoch ke-10, sedangkan penurunan kinerja pada epoch berikutnya menunjukkan adanya overfitting, yaitu keadaan di mana model secara konsisten menyelaraskan dirinya dengan data yang sedang dianalisis, sehingga membuatnya kurang ideal untuk data yang belum teramati. Jika dibandingkan dengan penelitian sebelumnya yang menggunakan BERT untuk domain yang dimaksud, seperti penelitian oleh Chantrapornchai & Tunasakul (2020) pada teks pariwisata dan Darji & Mitrovic (2023) pada dokumen hukum, penelitian ini menunjukkan bahwa model berbasis BERT memberikan kinerja yang tinggi ketika menggunakan korpus domain tertentu. Hal ini menunjukkan bahwa, sesuai dengan dataset domain dan konteks model, kemampuan model untuk memahami struktur data dan entitas yang muncul semakin meningkat. Namun, berbeda dengan penelitian ini, domain olahraga Indonesia memiliki karakteristik unik yaitu variasi entitas yang sangat banyak, seperti statistik pertandingan, skor, atau aksi atlet, sehingga tingkat kesulitan model relatif lebih tinggi.

Penggunaan metode rule-based untuk melakukan pemetaan entitas ke dalam format 5W1H dapat menampilkan hasil ekstraksi yang lebih teratur dan lebih jelas. Penggabungan model IndoBERT dan teknik rule-based dapat menghasilkan ekstraksi-ekstraksi yang lebih terstruktur yang lebih mudah dipahami oleh pengguna. Selain itu, penelitian ini memiliki beberapa keterbatasan. Pertama, kinerja model dipengaruhi oleh perbedaan jumlah data untuk setiap entitas, terutama untuk entitas yang sering muncul. Kedua, hanya tiga cabang olahraga (basket, bulu tangkis, dan sepak bola) yang termasuk dalam dataset, dan data dari ketiga media tersebut juga dimasukkan. Akibatnya, masih perlu dilakukan penelitian untuk menggeneralisasi model ke gaya penulisan olahraga atau media lain. Ketiga, kategori yang kompleks seperti STATISTIK dan AKSI menggunakan struktur numerik dan linguistik yang tidak biasa, seperti pola skor.

Secara keseluruhan, sistem ekstraksi informasi 5W1H yang menggunakan kombinasi IndoBERT dan berbasis rule-based dapat bekerja dengan baik dan memberikan hasil yang sesuai dengan tujuan. Ini menunjukkan bahwa model IndoBERT memiliki banyak potensi untuk diterapkan pada domain bahasa Indonesia, yang memungkinkan untuk meningkatkan proses 5W1H, meningkatkan domain, dan ukuran entitas. Hasil perancangan sistem ekstraksi data ini menunjukkan bahwa model IndoBERT dapat mendeteksi NER dengan cukup baik, dan metode berbasis aturan untuk pemetaan 5W1H telah dibuat dengan sukses. Model IndoBERT, yang telah disempurnakan dengan data berita olahraga Indonesia, dapat dengan mudah mengidentifikasi entitas olahraga. Hasil menunjukkan bahwa model menerima skor F1 terbaik sebesar 0,8191 pada epoch ke-10. Selain itu, penerapan algoritma berbasis aturan yang digunakan dalam desain ini menunjukkan kemampuan metode untuk menggabungkan hasil pengenalan entitas NER ke dalam format 5W1H yang terstruktur dan mudah dipahami. Metode ini memungkinkan sistem ekstraksi untuk menampilkan data penting dalam format 5W1H.

Untuk pengembangan lebih lanjut, desain ini masih memiliki banyak ruang yang besar untuk perbaikan. Salah satu contohnya adalah memungkinkan model ekstraksi ini menjadi lebih fleksibel dan bermanfaat untuk olahraga lain dengan entitas yang serupa dengan ketiga cabang olahraga (basket, bulutangkis dan sepak bola) dengan melatihnya dengan dataset olahraga lain dan sumber

media lainnya. Untuk meningkatkan akurasi pendeteksian entitas olahraga, juga dapat diterapkan pendekatan pembelajaran mendalam hibrida, seperti menggabungkan transformer dengan model pelabelan sekuens seperti CRF atau BiLSTM-CRF.

## UCAPAN TERIMAKASIH

Perancangan ini didukung oleh Ibu dan Bapak dosen pembimbing yang telah memberikan bimbingan,arahan, serta masukan yang sangat bermanfaat dan berharga selama peroses perancangan hingga penyusunan publikasi ini. Ucapan terimakasih juga disampaikan kepada keluarga terdekat serta teman-teman yang saya kasihi. Terimakasih ini disampaikan atas dukung yang diberikan untuk mendukung publikasi dari perancangan ini. Semua dukungan yang sudah diberikan merupakan bagian penting dalam menyelesaikan perancangan ini dan publikasi ini.

## DAFTAR PUSAKA

[1] APJII, "Jumlah Pengguna Internet Indonesia 2024," 7 2 2024. [Online]. Available: https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang.

[2] C. M. Annur, "Pengguna Internet di Indonesia Tembus 213 Juta Orang hingga Awal 2023," 20 09 2023. [Online]. Available: https://databoks.katadata.co.id/teknologi-telekomunikasi/statistik/d109a45f4409c34/pengguna-internet-di-indonesia-tembus-213-juta-orang-hingga-awal-2023.

[3] M. Indonesia, "55 Persen Waktu Masyarakat Indonesia Dihabiskan di Open Internet," 15 02 2023. [Online]. Available: https://mediaindonesia.com/humaniora/558604/55-persen-waktu-masyarakat-indonesia-dihabiskan-di-open-internet.

[4] D. Mardiah, "Minat Baca di Indonesia: Systematic Literature Review," *Jurnal Pena Ilmiah,* pp. 33-44, 2023.

[5] R. T. &. V. Gunawan, "Concentrated, Corporate, and Camouflaged: The Nature of AI News Coverage in Indonesia," *Asian Journal of Media and Communication,* vol. 8, no. 2, 23 12 2024.

[6] A. S. J. H. &. C. L. Jing Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* 2020.

[7] K. M.V, "BERT: A Review of Applications in Natural Language Processing and Understanding," *arXiv,* 2021.

[8] A. R. J. H. L. T. B. Fajri Koto, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *arXiv,* 2020.

[9] C. C. &. A. Tunsakul, "Information Extraction based on Named Entity for Tourism Corpus," *arXiv,* 2020.

[10] J. M. &. M. G. Harshil Darji, "German BERT Model for Legal Named Entity Recognition," *arXiv,* 2023.

[11] N. S. N. P. J. U. L. J. A. N. G. Ł. K. &. I. P. Ashish Vaswani, "Attention Is All You Need," *arXiv,* 2023.

[12] G. B. M. D. B. &. C. C. D. Tucudean, "Natural language processing with transformers: a review," *IEEE Access,* 2024.

[13] J. H. X. &. L. J. Gao, "A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of BERT," *IEEE,* 2020.

[14] B. V. K. W. G. I. C. S. L. X. L. Z. Y. S. S. M. R. F. P. B. S. &. P. A. Wilie, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing,* 2020.

[15] M. M. A. &. M. M. A. Buda, "A comprehensive survey of loss functions and metrics in deep learning," *Artificial Intelligence Review,* 2025.