nline pada laman web jurnal: **https://jurnal.plb.ac.id/index.php/tematik/index**

TEMATIK

*Jurnal* **Teknologi Informasi Komunikasi (e-Journal)**

Vol. 12 No. 1 (2025) 81 - 87 ISSN Media Elektronik: 2443-3640

# Ekstraksi Informasi dari Artikel Berita Agromaritim di Indonesia Menggunakan Teknik *Named Entity Recognition* (NER)

*Information Extraction from Agromatime Neswsarticles in Indonesia Using Named Entity Recognition (NER) Technique*

Adam Firmansyah Suhendar[1*], Wina Witanti[2], Melina[3]
[1,2,3]Prodi Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[1]adamfirmansyah479@gmail.com, [2]Witanti@gmail.com, [3]melina@lecture.unjani.ac. id

***Abstract***

*Indonesia as an archipelagic country has great potential in the agro-maritime sector, especially fisheries. However, geographic information from news articles is difficult to extract because the text is unstructured and there are language variations. This study aims to develop a geographic information extraction system from agro-maritime news using the Named Entity Recognition (NER) technique based on the BERT (Bidirectional Encoder Representations from Transformers) model. The BERT model was chosen because it is able to understand the context of words in depth, so it can recognize geographic entities such as the name of a water area even though there are variations in sentence structure. This study also produced a special annotated agro-maritime dataset used for model training. System evaluation was carried out using precision, recall, and F1-score metrics. The evaluation results showed that the model achieved a precision of 0.9768, recall of 0.9762, and F1-score of 0.9712, which indicated very good performance. The developed system is expected to improve the utilization of geographic information from news and become the basis for the development of data-based technology in the management of agro-maritime resources in Indonesia.*

*Keywords: agro-maritime news, BERT, geographical information, Named Entity Recognition, NLP.*

**Abstrak**

Indonesia sebagai negara kepulauan memiliki potensi besar di sektor agromaritim, khususnya perikanan. Namun, informasi geografis dari artikel berita sulit diekstraksi karena teks bersifat tidak terstruktur dan adanya variasi bahasa. Penelitian ini bertujuan mengembangkan sistem ekstraksi informasi geografis dari berita agromaritim dengan menggunakan teknik *Named Entity Recognition* (NER) berbasis model BERT *(Bidirectional Encoder Representations from Transformers)*. Model BERT dipilih karena mampu memahami konteks kata secara mendalam, sehingga dapat mengenali entitas geografis seperti nama wilayah perairan meskipun terdapat variasi struktur kalimat. Penelitian ini juga menghasilkan dataset agromaritim beranotasi khusus yang digunakan untuk pelatihan model. Evaluasi sistem dilakukan dengan metrik *precision*, *recall*, dan *F1-score*. Hasil evaluasi menunjukkan bahwa model mencapai *precision* 0,9768, *recall* 0,9762, dan *F1-score* 0,9712, yang mengindikasikan performa sangat baik. Sistem yang dikembangkan diharapkan dapat meningkatkan pemanfaatan informasi geografis dari berita serta menjadi dasar pengembangan teknologi berbasis data dalam pengelolaan sumber daya agromaritim di Indonesia.

Kata kunci: BERT, berita agromaritim, informasi geografis, *Named Entity Recognition*, NLP.

## 1. Pendahuluan

Indonesia memiliki potensi besar dalam sektor perikanan dan sumber daya alam lainnya. Potensi perikanan berkelanjutan Indonesia mencapai 12,54 juta ton per tahun [1]. Agromaritim Indonesia sangat relevan mengingat 70% wilayahnya adalah laut [2]. Namun, ekstraksi informasi geografis dari berita agromaritim menghadapi tantangan terkait variasi bahasa dalam teks yang sering mengakibatkan kesalahan pengenalan entitas [3].

Berita adalah sumber informasi penting, tetapi sering kali berbentuk teks tidak terstruktur, sehingga menyulitkan analisis data geografis. Metode *Natural Language Processing* (NLP) dapat digunakan untuk mengenali lokasi dan entitas geografis dalam berita, mempermudah pemetaan area perikanan dan identifikasi wilayah rawan bencana [4]. Salah satu model NLP yang kuat adalah BERT, yang dikembangkan oleh Google. BERT memahami konteks kata dengan melihat kata sebelum dan sesudahnya dalam kalimat. Model ini menggunakan mekanisme *attention* untuk menangkap hubungan antar kata secara menyeluruh [5]. Contoh mekanisme perhatian (*attention*) yang dihitung dapat dilihat pada Persamaan 1 [6]:

Diterima Redaksi: 15-05-2025 | Selesai Revisi: 06-06-2024 | Diterbitkan Online: 26-06-2025

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Q (*query*), K (*key*), dan V (*value*) adalah matriks hasil dari *input* yang diproses melalui *layer* linier. Hasil dari $QK^T$ dibagi dengan akar dari dimensi kunci (dk), lalu diproses dengan fungsi *softmax* untuk menentukan bobot perhatian, yang digunakan untuk menghasilkan representasi kontekstual kata.

Selama pelatihan, model menggunakan fungsi *loss* berbentuk *cross-entropy* seperti pada Persamaan 2 [6]:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (2)$$

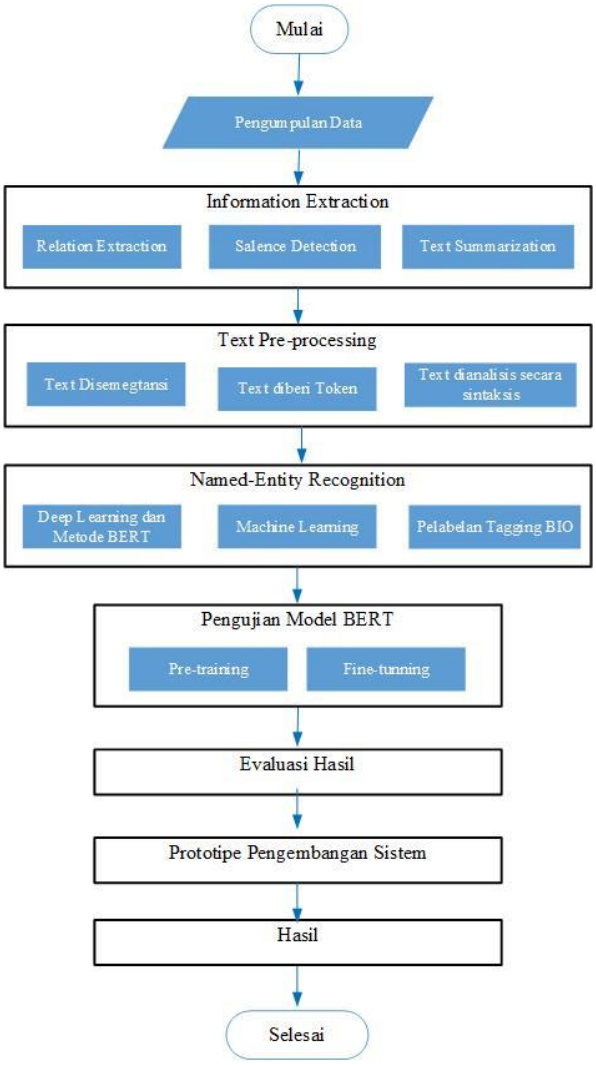
Model memanfaatkan fungsi loss yang dikenal sebagai *cross-entropy*, yang dihitung dengan mengalikan label kebenaran $y_i$ dengan logaritma dari probabilitas prediksi $(\hat{y}_i)$, kemudian menjumlahkannya untuk semua data. Fungsi ini mengukur selisih antara prediksi dan label sebenarnya, sehingga membantu model belajar mengenali entitas dengan lebih akurat.

Penelitian terdahulu menunjukkan bahwa BERT memiliki potensi besar dalam tugas NER, meskipun masih menghadapi tantangan dalam menangani variasi bahasa dan keterbatasan data[6]. Di bidang wisata, misalnya, pembuatan dataset khusus masih menjadi kendala karena keterbatasan sumber daya [7]. Penelitian lain menggabungkan NER dengan OCR untuk mengekstrak informasi dari poster digital, menghasilkan akurasi hingga 92% [8]. Pendekatan baru yang mengintegrasikan NER dengan model parsing ketergantungan juga menunjukkan peningkatan efisiensi dan akurasi [9]. Meskipun LSTM dan BiLSTM masih banyak digunakan, masih terbuka luas untuk pengembangan NER di sektor agromaritim, terutama melalui perluasan dataset dan penerapan *deep learning* yang lebih canggih [10].

Penelitian ini bertujuan mengembangkan sistem NER berbasis BERT untuk mengekstrak entitas geografis dari berita agromaritim di Indonesia. Sistem ini memanfaatkan kemampuan BERT dalam memahami konteks bahasa, sehingga mampu mengenali wilayah geografis secara otomatis meskipun terdapat variasi dalam struktur teks. Penelitian ini juga menghasilkan dataset beranotasi khusus untuk pelatihan model. Evaluasi sistem dilakukan menggunakan *F1-score*, *precision*, dan *recall*. Hasil akhirnya adalah prototipe sistem ekstraksi informasi geografis berbasis NLP, yang dapat membantu pengolahan dan analisis data menjadi lebih terstruktur dan aplikatif di berbagai bidang.

## 2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental yang berfokus pada pengembangan model NER dengan menggunakan algoritma BERT. Proses penelitian disusun dalam enam langkah utama seperti yang terlihat pada Gambar 1.




Gambar 1. Alur Penelitian

Tahap pertama yang dilakukan dalam penelitian ini adalah pengumpulan data yang dimulai dengan mencari sumber-sumber berita online yang terpercaya dan berkaitan, seperti portal berita nasional Kementerian Kelautan dan Perikanan (KKP), serta situs berita khusus yang mengupas isu-isu agromaritim. Pengumpulan data dilakukan secara terstruktur menggunakan teknik *web scraping* atau pengunduhan manual, bergantung pada struktur dan kebijakan setiap situs web. Target jumlah artikel yang dikumpulkan adalah 2500 artikel, dengan rentang waktu penerbitan yang disesuaikan untuk memastikan keberagaman dan representasi data yang maksimal.

Setelah data terkumpul, dilakukan ekstraksi informasi melalui tiga metode utama, yaitu *Relation Extraction*, *Salience Detection*, dan *Text Summarization* [11]. *Relation Extraction* dilakukan menggunakan pendekatan berbasis aturan dan dependensi sintaksis, untuk mengidentifikasi relasi antar entitas dalam satu kalimat [12]. *Salience Detection* diterapkan untuk

DOI: https://doi.org/10.38204/tematik.v12i1.2329

menyaring pernyataan atau entitas yang memiliki bobot informasi signifikan menggunakan aturan (*rule-based method*), untuk mengidentifikasi kalimat-kalimat yang paling menonjol atau informatif dalam sebuah teks. Selanjutnya, dilakukan *Text Summarization* menggunakan pendekatan ekstraktif berbasis GraphRank untuk merangkum bagian penting dalam artikel [13].

Proses berikutnya adalah *Text Pre-processing*, yang mencakup segmentasi kalimat dengan pustaka nltk [14]. Tokenisasi menggunakan BERT tokenizer, dan analisis sintaksis dengan spaCy versi bahasa Indonesia. Tahap ini penting untuk menyiapkan data yang akan diberi anotasi [15]. Entitas yang ditandai meliputi nama wilayah geografis, wilayah administratif, sungai, laut, dan lokasi spesifik lain. Proses pelabelan dilakukan dengan skema *Begin*, *Inside*, *Outside* (BIO), dan menghasilkan data dalam format standar CoNLL yang umum digunakan dalam pelatihan model NER [16].

Pengembangan model NER dilakukan dengan menggunakan model *pre-trained* BERT yang dilatih ulang secara khusus untuk mengenali entitas geografis dalam teks berita [6]. Data pelatihan berupa teks berita yang telah dianotasi secara otomatis dengan skema BIO untuk menandai entitas seperti nama wilayah administratif, kawasan perairan, dan lokasi spesifik lainnya. Dataset terdiri dari 2.500 artikel berita yang telah diproses menjadi format CoNLL, dengan total 35.000 token [17].

Pelatihan dilakukan menggunakan arsitektur Transformer BERT berbasis *Deep Neural Network* melalui API Trainer dari pustaka Hugging Face Transformers. Data pelatihan dibagi menjadi dua subset: 80% untuk pelatihan dan 20% untuk validasi. Proses pelatihan berlangsung selama 5 *epoch* dengan pengaturan hyperparameter: *batch size* = 16, *learning rate* = 5e-5, dan *maximum sequence length* = 128 token. Optimizer yang digunakan adalah AdamW, serta diterapkan *learning rate scheduler linier* untuk stabilisasi proses pembelajaran [18]. Replikasi hasil dimungkinkan dengan menyertakan random seed tetap dan *checkpoint* pelatihan yang disimpan setiap *epoch* [19].

Pengujian model dilakukan menggunakan dataset uji yang terdiri dari 500 artikel berita yang tidak termasuk dalam data latih maupun *pre-trained* corpus, guna menjamin objektivitas hasil evaluasi. Evaluasi dilakukan dengan pendekatan *fine-tuning*, di mana model diuji pada data baru dan hasilnya dibandingkan dengan label anotasi asli. Metrik yang digunakan untuk mengukur performa model meliputi *precision*, r*ecall*, dan F1-s*core*, yang dihitung untuk tiap label entitas serta secara keseluruhan menggunakan rata-rata mikro dan makro [20].

Evaluasi juga mencakup analisis kesalahan menggunakan confusion matrix dan distribusi per kelas entitas untuk mengidentifikasi kelemahan model. Model dinyatakan berhasil apabila mencapai skor F1 di atas 85% pada entitas utama. Semua proses evaluasi dilakukan dalam lingkungan Python 3.11 menggunakan *scikit-learn* dan pustaka seqeval sebagai pelengkap untuk metrik NER [5].

## 3. Hasil dan Pembahasan

Penelitian ini dimulai dengan pengumpulan data yang dilakukan secara terstruktur dari sejumlah sumber berita daring yang dapat dipercaya mengenai topik agromaritim, termasuk portal Kementerian Kelautan dan Perikanan (KKP) dan beberapa media nasional. Sebanyak 2.500 tulisan berhasil dihimpun dalam kurun waktu 2021 hingga 2024. Pengumpulan dilakukan melalui kombinasi metode web scraping dan pengunduhan manual, tergantung pada struktur HTML serta izin akses dari setiap situs web. Artikel kemudian dikonversi dan disimpan dalam format Excel untuk memudahkan pengolahan data lanjutan. Pada statistik pengumpulan data dapat dilihat pada Tabel 1.

Tabel 1. Statistik Pengumpulan Data

| Komponen Data | Nilai |
|---|---|
| Sumber Artikel | KKP, Media Nasional |
| Metode Pengumpulan | Web scraping, unduh manual |
| Jumlah Artikel | 2.500 |
| Rentang Waktu | 2021-2024 |
| Format Akhir | Excel (.xlsx) |

Setelah pengumpulan data selesai, ekstraksi informasi dilakukan melalui tiga metode: *relation extraction*, salience detection, dan *text summarization*. Pada tahap ekstraksi relasi, pendekatan berbasis aturan dan sintaksis dependensi digunakan untuk mengenali hubungan antar entitas seperti Lokasi–Kegiatan atau Wilayah–Komoditas. Metode ini dapat mengidentifikasi pola yang berulang dalam artikel yang mengaitkan daerah geografis dengan aktivitas atau jenis komoditas perikanan.

*Deteksi salience* dilakukan dengan metode heuristik yang menggabungkan jumlah entitas bernama dalam sebuah kalimat dan panjang kalimat tersebut (jumlah kata dibagi 10) untuk menetapkan nilai kepentingan (*salience score*). Kalimat-kalimat yang memiliki skor tertinggi dinilai paling informatif dan dipilih untuk analisis lebih lanjut. Metode ini menghasilkan sekitar 2.300 kalimat yang diidentifikasi sebagai kalimat krusial dalam seluruh korpus berita. Penerapan ringkasan teks dilakukan dengan metode ekstraktif GraphRank untuk merangkum inti artikel, menghasilkan ringkasan yang mencakup 20–30% dari isi teks aslinya. Contoh hasil dari metode GraphRank ekstraktif yang bisa merangkum informasi berita menjadi hanya berisi kata-kata penting dari isi berita tersebut dapat dilihat pada Tabel 2.

DOI: https://doi.org/10.38204/tematik.v12i1.2329

Seluruh data artikel yang sudah diringkas dan diklasifikasikan selanjutnya diproses dalam tahap pra-pemrosesan teks. Proses ini melibatkan pemecahan kalimat dengan pustaka NLTK, tokenisasi menggunakan tokenizer BERT, dan analisis sintaksis melalui spaCy untuk Bahasa Indonesia. Data selanjutnya dianotasi dengan skema pelabelan BIO (*Begin, Inside, Outside*) dan diubah ke dalam format standar CoNLL. Jumlah token yang diperoleh dari seluruh dataset mencapai sekitar 35.000 token, yang mencakup berbagai entitas seperti nama-nama laut, sungai, provinsi, dan daerah administratif lainnya. Contoh hasil pelabelan BIO yang sudah dilakukan dalam pengenalan entitas pada informasi berita perikanan dapat dilihat pada Tabel 3.

Tabel 2. Contoh Hasil Informasi Ekstraksi Berita Perikanan

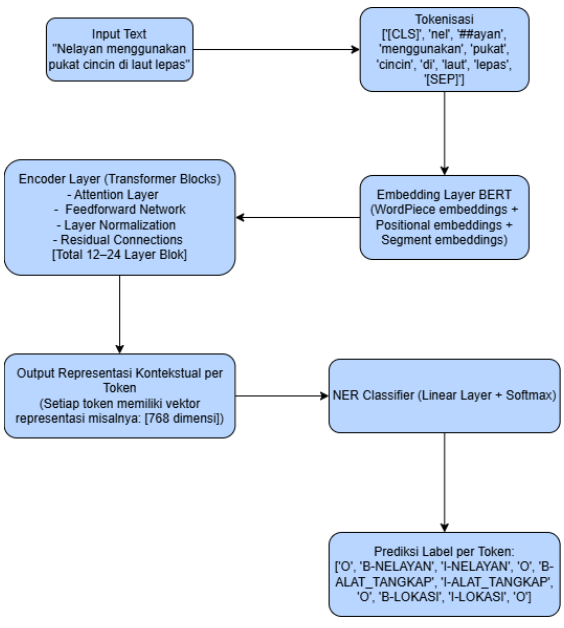
| Attribute | Data |
|---|---|
| Id | 1 |
| Content | pembangunan sektor perikanan diartikan upaya memanfaatkan sumber daya hayati perikanan sumber daya perairan kegiatan penangkapan budidaya seiring pengembangan penerapan iptek pengembangan produk peningkatan pendapatan kesejahteraan peningkatan kesempatan kerja berusaha peningkatan devisa negara disertai upaya pemeliharaan pelestarian sumber daya hayati lingkungan lestari KKP pembangunan perikanan menghilangkan kemiskinan pengangguran basis ekonomi nasional memiliki peluang dikembangkan sektor unggulan perdagangan internasional kurniawan Bappenas isu pengelolaan perikanan tangkap *illegal unregulated unreported* iuu fishing adalah rendahnya produktivitas daya saing teori pertumbuhan neo klasik dikemukakan oleh solow dan mankiw output perekonomian salah satu faktor pembangunan modal |
| Summary | Pembangunan sektor perikanan merupakan upaya pemanfaatan sumber daya hayati perairan melalui penangkapan dan budidaya, disertai penerapan iptek dan pengembangan produk untuk meningkatkan pendapatan, kesejahteraan, dan devisa negara. Upaya ini juga mencakup pelestarian lingkungan dan pengentasan kemiskinan. Namun, sektor ini menghadapi tantangan seperti IUU fishing, rendahnya produktivitas, dan daya saing, yang memerlukan pendekatan pembangunan berbasis teori pertumbuhan ekonomi klasik seperti Solow dan Mankiw. |

Tabel 3. Contoh Hasil Pelabelan Bio pada Entitas Perikanan

| Token | Tag_Bio |
|---|---|
| nelayan | B-NELAYAN |
| menggunakan | O |
| pukat | B-ALAT_TANGKAP |
| cincin | I-ALAT_TANGKAP |
| untuk | O |
| menangkap | O |
| ikan | O |
| di | O |
| laut | O |
| lepas | O |

Model pengenalan entitas NER selanjutnya dikembangkan dengan memakai arsitektur BERT. Model dasar BERT dijadikan sebagai dasar dan dilakukan pelatihan ulang (*fine-tuning*) menggunakan data yang telah dianotasi. Pelatihan dilaksanakan selama 5 *epoch* dengan memanfaatkan API Trainer dari pustaka Hugging Face Transformers, menggunakan parameter: ukuran batch 16, laju pembelajaran 5e-5, dan panjang urutan maksimum 128 token. Dataset terbagi menjadi 80% untuk data pelatihan (2.000 artikel) dan 20% untuk data validasi (500 artikel). Untuk menjaga stabilitas pelatihan, digunakan optimizer AdamW serta penjadwal *learning rate linier*. Model disimpan sebagai *checkpoint* di setiap epoch sehingga eksperimen dapat direplikasi dengan konsisten.

Alur identifikasi entitas seperti "nelayan" diidentifikasi menggunakan skema BIO berdasarkan konteks kalimat yang dievaluasi oleh model BERT dapat dilihat pada Gambar 2.




Gambar 2. Diagram Arsitektur BERT untuk NER

Diagram Arsitektur BERT untuk tugas NER, dimulai dari input teks "Nelayan menggunakan pukat cincin di laut lepas" yang menjalani proses tokenisasi dengan

DOI: https://doi.org/10.38204/tematik.v12i1.2329

teknik WordPiece, menghasilkan token seperti `[CLS]`, `nel`, `##ayan`, dan `[SEP]`. Token-token tersebut selanjutnya diolah di *Embedding Layer* yang mengintegrasikan WordPiece embedding, *positional embedding*, dan *segment embedding* untuk menciptakan representasi awal dari tiap token. Selanjutnya, representasi ini dikirimkan ke *Encoder Layer* yang terdiri dari beberapa blok transformer, yang mencakup *attention layer, feedforward network, layer normalization*, dan *residual connections*, untuk memahami konteks kata secara dua arah dalam kalimat. *Output* dari encoder merupakan representasi kontekstual setiap token dalam bentuk vektor berdimensi tinggi (contohnya 768 dimensi), yang selanjutnya diteruskan ke NER *Classifier* berupa lapisan linear dan softmax untuk mengaitkan setiap token dengan label entitas seperti `B-NELAYAN`, `I-ALAT_TANGKAP`, atau `B-LOKASI`. Struktur ini memungkinkan BERT dapat mengidentifikasi entitas dalam teks dengan lebih tepat meskipun ada variasi dalam susunan kalimat dan penggunaan kata yang rumit.

Prediksi mengindikasikan cara model mengidentifikasi entitas dalam kalimat panjang yang berhubungan dengan industri perikanan. Dalam kalimat itu, model mampu mengidentifikasi entitas "nelayan" meskipun kata tersebut terbagi menjadi dua token sub-kata oleh tokenizer BERT, yakni "nel" dan "##ayan". Kedua token ini diidentifikasi sebagai komponen dari entitas NELAYAN dengan skor kepercayaan yang sangat tinggi, yaitu 0.9875 dan 0.9859, yang menunjukkan bahwa model sangat percaya pada prediksi tersebut.

Hasil prediksi yang menyajikan informasi penting seperti indeks sampel, teks sumber, token yang dikenali sebagai bagian dari entitas, label entitas (dalam hal ini NELAYAN), dan skor kepercayaan untuk setiap token. Ini menunjukkan bahwa model dapat menangani tokenisasi sub-kata dengan baik dan menyatukan kembali untuk menghasilkan prediksi entitas yang tepat dapat dilihat pada Tabel 4.

Setelah pelatihan selesai, model diuji dengan menggunakan dataset uji yang berisi 500 artikel baru. Evaluasi dilakukan dengan menghitung metrik kunci dalam sistem NER yaitu: presisi, *recall*, dan F1-*score*. Hasil pengujian menunjukkan bahwa model memperoleh p*recision* 0,9768, *recall* 0,9762, dan F1-*score* 0,9712, yang mengindikasikan bahwa model tersebut berperforma sangat baik dan melampaui batas minimum keberhasilan (F1 > 85%). Evaluasi dilakukan dengan memanfaatkan pustaka *scikit-learn* dan seqeval dalam lingkungan Python 3.11. Hasil evaluasi model NER dengan menggunakan BERT dapat dilihat pada Tabel 5.

Hasil prediksi logit terbaru ini menunjukkan bahwa model secara konsisten memberikan nilai tertinggi pada kelas O (*Other*) atau I-LOC untuk sebagian besar token yang diberi label I-LOC, mencerminkan tingkat keakuratan yang tinggi dalam mengidentifikasi entitas geografis yang terdapat dalam teks berita. Analisis ini menegaskan bahwa model tidak hanya menunjukkan performa evaluasi kuantitatif yang tinggi (F1: 0.9712), tetapi juga dapat menghasilkan *output* logit yang logis dan sesuai dengan ekspektasi distribusi. Model cenderung konsisten dalam memberi prioritas pada kelas yang tepat di antara lima kemungkinan label (BIO+O). Hasil Logit Prediksi dan Label dapat dilihat pada Tabel 6.

Tabel 4. Contoh Hasil Pelatihan Model NER Menggunakan BERT

| Sample | Text | Entity | Label | Score |
|---|---|---|---|---|
| 0.1 | industri perikanan industri penangkapan ikan industri aktivitas menangkap membudi dayakan memproses mengawetkan menyimpan mendistribusikan memasarkan produk ikan istilah didefinisikan fao mencakup pemancing rekreasi nelayan tradisional penangkapan ikan | nel | NELAYAN | 0.9875 |
| 1.1 | industri perikanan industri penangkapan ikan industri aktivitas menangkap membudi dayakan memproses mengawetkan menyimpan mendistribusikan memasarkan produk ikan istilah didefinisikan fao mencakup pemancing rekreasi nelayan tradisional penangkapan ikan | ##ayan | NELAYAN | 0.9859 |

Tabel 5 Hasil Evaluasi Model NER Menggunakan BERT

| Metrik Evaluasi | Nilai |
|---|---|
| Precision | 0.9768 |
| Recall | 0.9762 |
| F1-Score | 0.9712 |
| Test_Loss | 0.2111 |
| Kecepatan | 1.831 |

DOI: https://doi.org/10.38204/tematik.v12i1.2329

Tabel 6. Hasil Prediksi Logit dan Label

| Token ke- | Logits per Kelas (B-LOC, I-LOC, B-ORG, I-ORG, O) | Label_ID |
|---|---|---|
| 1 | [-1.3916, -1.1290, -1.6729, -1.2997, 4.5051] | -100 |
| 2 | [-2.2939, -1.9844, -2.5584, -1.5704, 6.8192] | 4 |
| 3 | [-2.3524, -2.0589, -2.5881, -1.8020, 7.0086] | 4 |
| 4 | [-2.3881, -2.1128, -2.5616, -1.7190, 7.0010] | 4 |
| 5 | [-2.3937, -2.0672, -2.5854, -1.6819, 6.9878] | 4 |

Dalam proses analisis logit, hasil prediksi model masih menghadapi beberapa kendala yang muncul. Salah satu tantangan utama adalah adanya kesamaan skor logit di antara kelas negatif, seperti B-LOC, I-LOC, B-ORG, dan I-ORG, yang semua memiliki nilai rendah dan berdekatan satu sama lain, sedangkan kelas O (*Other*) selalu jauh lebih tinggi. Akibatnya, model ini sangat condong pada prediksi kelas O, meskipun token seharusnya termasuk dalam kelas entitas seperti I-LOC. Situasi ini mengindikasikan bahwa ketidakseimbangan distribusi label dalam data pelatihan (di mana sebagian besar token adalah O) membuat model kesulitan dalam membedakan dengan tepat antara entitas khusus, terutama pada dokumen panjang atau entitas yang memiliki konteks berdekatan. Selain itu, walaupun nilai logit memperlihatkan kecenderungan yang tepat, model masih memerlukan metode pasca-pelatihan seperti penyesuaian bobot kelas atau peningkatan data untuk memperbaiki akurasi pada entitas yang minor.

Model BERT yang digunakan dalam penelitian ini menunjukkan kinerja yang baik secara kuantitatif, tetapi masih memiliki beberapa keterbatasan yang perlu diperhatikan. Salah satu tantangan utamanya terletak pada analisis logit, di mana model cenderung menghasilkan skor tertinggi untuk kelas O (*Other*), sementara kelas-kelas entitas lainnya seperti B-LOC, I-LOC, B-ORG, dan I-ORG memiliki skor logit yang rendah dan saling berdekatan. Ketidakseimbangan ini membuat model lebih sering mengklasifikasikan token sebagai O, meskipun secara konteks token tersebut seharusnya masuk dalam kategori entitas.

Fenomena ini menunjukkan adanya ketidakseimbangan distribusi label dalam data pelatihan, di mana sebagian besar token tidak tergolong sebagai entitas bernama dan diberi label O. Ketidakseimbangan ini menyulitkan model untuk membedakan dan mengenali entitas yang signifikan, terutama pada dokumen panjang atau saat entitas berdekatan dalam satu kalimat.

Di samping itu, model menghadapi tantangan dalam mengatasi entitas yang saling tumpang tindih (*nested entities*), yaitu kondisi di mana ada dua atau lebih entitas yang tumpang tindih dalam satu frasa. Metode pelabelan BIO yang diterapkan dalam penelitian ini tidak dapat menangani kompleksitas tersebut dengan baik. Selain itu, entitas minor seperti nama tempat lokal atau institusi kecil yang jarang ada dalam data pelatihan biasanya diabaikan oleh model karena kurang representatif, sehingga menurunkan akurasi deteksi terhadap entitas-entitas ini. Lebih lanjut, model juga menghadapi tantangan dalam melakukan disambiguasi entitas, yaitu kemampuan untuk membedakan makna dari suatu entitas yang memiliki banyak kemungkinan referensi tergantung pada konteks. Entitas geografis di Indonesia, misalnya, sering kali memiliki nama yang ambigu. Kata “Aceh” dapat merujuk pada provinsi, wilayah konflik dalam konteks sejarah, atau kawasan yang terkena dampak tsunami. Demikian pula, “Laut Jawa” bisa merujuk pada lokasi geografis formal atau wilayah penangkapan ikan tertentu. Jika konteks tidak dipahami secara utuh oleh model, maka kemungkinan kesalahan klasifikasi menjadi tinggi.

Untuk mengatasi berbagai keterbatasan tersebut, diperlukan strategi pasca-pelatihan seperti penyesuaian bobot kelas *(class weight adjustment)*, peningkatan proporsi entitas minor dalam dataset, serta eksplorasi metode lanjutan seperti *Conditional Random Fields* (CRF) atau pendekatan *multi-task learning* yang mampu menangani struktur entitas kompleks dan meningkatkan kemampuan disambiguasi yang lebih kuat. Strategi-strategi ini diharapkan dapat meningkatkan performa model secara menyeluruh, baik dari sisi ketepatan klasifikasi maupun pemahaman konteks semantik.

## 4. Kesimpulan

Berdasarkan temuan penelitian yang telah dijelaskan sebelumnya, dapat disimpulkan bahwa penelitian ini berhasil mengembangkan sistem untuk mengekstraksi entitas geografis dari berita agromaritim di Indonesia dengan menggunakan teknik *Named Entity Recognition* (NER) yang berbasis model BERT. Sistem yang dirancang dapat secara otomatis mengidentifikasi area administratif, kawasan perairan, sungai, dan lokasi geografis lainnya dengan kinerja yang unggul, terbukti oleh nilai *precision* 0,9768, *recall* 0,9762, dan *F1-score* 0,9712. Metode pelabelan dengan skema BIO dan format CoNLL terbukti berhasil dalam melatih model dan menghasilkan *output* yang tepat. Sistem ini memiliki potensi yang signifikan untuk diterapkan dalam analisis spasial yang berbasis teks, membantu penyusunan kebijakan kelautan, serta memantau wilayah agromaritim dengan cara yang dinamis dan efektif.

Meskipun kinerja model tergolong baik, penelitian ini masih menghadapi sejumlah tantangan, khususnya dalam klasifikasi entitas yang serupa, penanganan entitas yang saling tumpang tindih, serta disambiguasi entitas dengan makna ganda sesuai konteks. Oleh sebab itu, pengembangan sistem di masa mendatang harus meliputi pendekatan disambiguasi yang lebih efektif,

DOI: https://doi.org/10.38204/tematik.v12i1.2329

peningkatan ragam entitas yang dikenali, serta integrasi sistem ke dalam platform berbasis web atau GIS yang lebih interaktif.

Sebagai rekomendasi, sistem ini memiliki peluang signifikan untuk diimplementasikan di industri, khususnya dalam sektor kelautan, perikanan, dan logistik berbasis wilayah. Instansi pemerintah maupun perusahaan maritim, misalnya, dapat memanfaatkan sistem ini untuk mengotomatiskan pemetaan informasi geografis dari laporan berita dan media sosial guna mendukung pengambilan keputusan cepat dalam pengelolaan wilayah laut, prediksi bencana kelautan, serta pemantauan aktivitas penangkapan ikan ilegal. Selain itu, perusahaan teknologi yang menciptakan sistem pemrosesan berita atau pelacakan data geografis juga bisa menyesuaikan model ini untuk memperbaiki ketepatan ekstraksi lokasi secara otomatis.

Untuk penelitian selanjutnya, disarankan agar jenis entitas yang dikenali diperluas, mencakup entitas non-geografis yang relevan dalam konteks agromaritim, seperti jenis komoditas, alat tangkap, dan pelaku usaha. Penelitian lebih lanjut juga direkomendasikan untuk menguji kinerja sistem pada teks dalam berbagai bahasa daerah atau multilingual, untuk meningkatkan generalisasi model di berbagai konteks lokal. Selain itu, pengembangan antarmuka pengguna dan integrasi dengan sistem informasi geografis dapat menjadi langkah strategis untuk memastikan sistem ini dapat digunakan secara luas dalam industri dan pemerintahan.

## Daftar Rujukan

[1] U. A. Nugroho and F. Budianto, "Perspektif Eksploitasi dan Konservasi dalam Pengelolaan Sumber Daya Perikanan Indonesia," *J. Media Perencana*, vol. 2, no. 1, pp. 51–67, 2021,[Online].Available:https://mediaperencana.perencanapembangunan.or.id/index.php/mmp/article/view/20/13

[2] M. A. Fitri and M. Usni, "Strategi Pengembangan Agro Maritim Di Wilayah Pesisir Kota Padang Sumatera Barat," *Semin. Nas. Has. Penelit. Kelaut. dan Perikan.* , vol. 5587, pp. 90–95, 2022.

[3] N. Nurwanda, N. Suarna, and W. Prihartono, "Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1841–1846, 2024, doi: 10.36040/jati.v8i2.8469.

[4] E. C. Sukmawati, L. Suryaningrum, and D. Angelica, "SisInfo Klasifikasi Berita Palsu Menggunakan Model Bidirectional Encoder Representations From Transformers ( BERT ) SisInfo," vol. 6, no. 2, pp. 76–85, 2024.

[5] A. R. Hanum *et al.*, "Mendeteksi Berita Hoaks Performance Analysis of the Bert Text Classification Algorithm," vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik938093.

[6] S. Naseer *et al.*, "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance.," *Pakistan J. Multidiscip. Res.*, vol. 2, no. 2, pp. 293–308, 2021.

[7] A. Zahra, A. F. Hidayatullah, and S. Rani, "Kajian Literatur Named Entity Recognition pada Domain Wisata," *Automata*, vol. 2, no. 1, pp. 0–4, 2021.

[8] A. S. Rosidy, T. M. Akhriza, and M. Husni, "Combining the NER-OCR methods to improve information retrieval efficiency in the Indonesian posters," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, pp. 263–269, 2020, doi: 10.14710/jtsiskom.2020.13686.

[9] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 6470–6476, 2020, doi: 10.18653/v1/2020.acl-main.577.

[10] M. Fakhri, D. A. Putra, and A. Fathan Hidayatullah, "Tinjauan Literatur : Named Entity Recognition pada Ulasan Wisata," *Automata*, vol. 2, no. 1, 2021.

[11] M. R. Azizi, W. Hayuhardhika, N. Putra, and I. Arwani, "Ekstraksi Informasi pada Data Logbook KKN Mahasiswa Fakultas Ilmu Komputer Universitas Brawijaya Malang menggunakan Metode NER (Named Entity Recognition)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, pp. 2895–2903, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[12] K. Detroja, C. K. Bhensdadia, and B. S. Bhatt, "A survey on Relation Extraction," *Intell. Syst. with Appl.*, vol. 19, no. July2022,p.200244,2023,doi: 10.1016/j.iswa.2023.200244.

[13] R. Hadwirianto, "Extractive Text Summarization Terhadap Artikel Berita Indonesia Berbasis Machine Learning," vol. 11, no. 4, pp. 3941–3946, 2024.

[14] C. Magdalena and B. H. Tambun, "Segmentasi Dokumen Teks Dengan Metode Texttiling," *J. Ilm. Inform.*, vol. 10, no. 01, pp. 8–14, 2022, doi: 10.33884/jif.v10i01.4509.

[15] F. Lubis *et al.*, "Penggunaan Metode Text Mining Untuk Mengekstrak Informasi Penting Dari Teks Laporan Penelitian," *J. Motiv. Pendidik. dan Bhs.*, vol. 1, no. 4, 2023, [Online]. Available: https://doi.org/10.59581/jmpb-widyakarya.v1i4.1961

[16] A. D. Ariyadi and A. P. Y. Utomo, "Analisis Kesalahan Sintaksis pada Teks Berita Daring berjudul Mencari Etika Elite Politik di saat Covid-19," *J. Bhs. dan Sastra*, vol. 8, no. 3, p. 138, 2020, doi: 10.24036/jbs.v8i3.110903.

[17] W. Pinasti and L. H. Suadaa, "Named Entity Recognition pada Kueri Pencarian Statistik," vol. 13, pp. 171–177, 2024.

[18] A. A. Mudding, "Mengungkap Opini Publik: Pendekatan BERT-based-caused untuk Analisis Sentimen pada Komentar Film," *J. Syst. Comput. Eng.*, vol. 5, no. 1, pp. 36–43, 2024, doi: 10.61628/jsce.v5i1.1060.

[19] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting Few-Sample Bert Fine-Tuning," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, pp. 1–22, 2021.

[20] D. T. Arum and A. I. Pradana, "Implementasi Bidirectional Encoder Representations From Transformers (BERT) Untuk Klasifikasi Spam Pada Email," vol. 9, no. 2, pp. 2491–2496, 2025.

DOI: https://doi.org/10.38204/tematik.v12i1.2329