Laporan Project Akhir Pemrosesan Bahasa Alami

EKSTRAKSI ENTITAS BISNIS DARI ARTIKEL KEUANGAN INDONESIA MENGGUNAKAN FINE-TUNING INDOBERT

Disusun untuk memenuhi

Tugas Mata Kuliah Pemrosesan Bahasa Alami

Oleh:

AHMAD SYAH RAMADHAN (2208107010033)

WIDYA NURUL SUKMA (2208107010054)

JURUSAN INFORMATIKA
FAKULTAS METEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
2025

DAFTAR ISI

DAFTAR ISI **2**
BAB I PENDAHULUAN **4**
1.1 Latar Belakang 4
1.2 Rumusan Masalah 4
1.3 Tujuan 4
1.4 Batasan Masalah 4
1.5 Manfaat 5
BAB II KAJIAN PUSTAKA **6**
2.1 Natural Language Processing 6
2.2 Named Entity Recognition (NER) 6
2.3 Model Transformer dan BERT 6
2.4 IndoBERT dan Pre-trained Model untuk Bahasa Indonesia 6
2.5 Fine-tuning Model Pre-trained untuk NER 6
BAB 3 METODOLOGI **7**
3.1 Desain Project 7
3.2 Dataset 7
3.2.1 Pengumpulan Data 7
3.2.2 Anotasi Data 7
3.2.3 Pembagian Dataset 8
3.3 Preprocessing Data 8
3.4 Arsitektur Model 8
3.4.1 IndoBERT 8
3.4.2 IndoRoBERTa 8
3.4.3 Token Classification Head 9
3.5 Proses Fine-tuning 9
3.5.1 Hyperparameter 9
3.5.2 Strategi Training 9
3.6 Evaluasi Model 9
3.7 Implementasi Aplikasi Web 10
BAB 4 HASIL DAN PEMBAHASAN **11**
4.1 Statistik Dataset 11
4.2 Hasil Fine-tuning 11
4.2.1 Training Loss dan Validation Loss 11
4.2.2 Performa Model IndoBERT 12
4.2.3 Performa Model IndoRoBERTa 12
4.3 Perbandingan Model 12

4.4 Aplikasi Web 13
BAB 5 KESIMPULAN **16**
DAFTAR PUSTAKA **17**

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang berkembang pesat, informasi keuangan menjadi salah satu aspek penting dalam pengambilan keputusan bisnis dan ekonomi. Artikel-artikel keuangan dari media online banyak memuat informasi strategis seperti nama perusahaan, angka finansial, dan tokoh-tokoh eksekutif yang relevan. Namun, mengekstrak informasi ini secara manual membutuhkan waktu, tenaga, dan biaya yang besar, terutama dengan volume data yang terus meningkat setiap harinya.

Teknologi Natural Language Processing (NLP), khususnya Named Entity Recognition (NER), menawarkan solusi otomatis untuk mengekstrak entitas penting dari artikel secara efisien. Dalam konteks Bahasa Indonesia, model pre-trained berbasis transformer seperti IndoBERT yang dirancang khusus untuk memahami struktur dan kosa kata Bahasa Indonesia memiliki potensi besar dalam meningkatkan akurasi ekstraksi entitas dari artikel keuangan. Dengan memanfaatkan IndoBERT untuk tugas NER, proses analisis informasi keuangan dapat dilakukan lebih cepat dan hemat sumber daya.

1.2 Rumusan Masalah

1. Bagaimana mengembangkan sistem NER berbasis IndoBERT untuk mengekstrak entitas bisnis dari artikel keuangan Indonesia?
2. Bagaimana perbandingan performa antara model IndoBERT dan IndoRoBERTa dalam tugas NER untuk artikel keuangan?
3. Bagaimana implementasi model NER dalam aplikasi web (streamlit) yang dapat digunakan untuk analisis artikel keuangan?

1.3 Tujuan

1. Mengembangkan sistem NER berbasis IndoBERT untuk mengekstrak entitas bisnis dari artikel keuangan Indonesia.
2. Membandingkan performa IndoBERT dengan IndoRoBERTa dalam tugas ekstraksi entitas bisnis.
3. Membangun aplikasi web menggunakan Streamlit untuk ekstraksi entitas dari artikel keuangan.

1.4 Batasan Masalah

1. Penelitian fokus pada ekstraksi tiga jenis entitas: organisasi (perusahaan/lembaga), angka finansial, dan tokoh eksekutif.

2. Dataset dari portal berita keuangan Indonesia (sumber: kontan.co.id dan bisnis.com) dengan 2.200 artikel.
3. Model yang digunakan terbatas pada IndoBERT dan IndoRoBERTa sebagai pembanding.

1.5 Manfaat

1. Kontribusi pada pengembangan teknologi NLP untuk Bahasa Indonesia dalam domain keuangan.
2. Memfasilitasi analisis data keuangan lebih efisien untuk riset, jurnalisme, dan pengambilan keputusan bisnis.
3. Menyediakan pondasi untuk pengembangan sistem analisis sentimen dan pemetaan relasi entitas bisnis di Indonesia.

BAB 2
KAJIAN PUSTAKA

2.1 Natural Language Processing

Natural Language Processing (NLP) adalah cabang ilmu komputer yang fokus pada interaksi antara komputer dan bahasa manusia. Perkembangan NLP telah melalui beberapa tahap, dari pendekatan berbasis aturan hingga metode berbasis statistik dan deep learning.

2.2 Named Entity Recognition (NER)

Named Entity Recognition adalah tugas mengidentifikasi dan mengklasifikasikan entitas bernama dalam teks ke dalam kategori yang telah ditentukan. Dalam konteks keuangan, NER memiliki karakteristik dan tantangan khusus.

2.3 Model Transformer dan BERT

Transformer adalah arsitektur neural network yang mengandalkan mekanisme self-attention. BERT (Bidirectional Encoder Representations from Transformers) merupakan implementasi transformer yang menghasilkan representasi kontekstual bidirectional.

2.4 IndoBERT dan Pre-trained Model untuk Bahasa Indonesia

IndoBERT adalah model BERT yang dilatih khusus untuk Bahasa Indonesia. Model ini telah menunjukkan performa yang baik dalam berbagai tugas NLP untuk Bahasa Indonesia.

2.5 Fine-tuning Model Pre-trained untuk NER

Fine-tuning adalah proses melatih model pre-trained pada dataset spesifik untuk tugas tertentu, dalam hal ini NER untuk domain keuangan.

BAB 3
METODOLOGI

3.1 Desain Project

Project ini menggunakan pendekatan eksperimental dengan desain komparatif untuk mengembangkan dan mengevaluasi sistem Named Entity Recognition (NER) berbasis transformer untuk ekstraksi entitas bisnis dari artikel keuangan Indonesia. Pendekatan yang digunakan meliputi:

1. **Tahap Persiapan Data**: Pengumpulan dan anotasi dataset artikel keuangan dari portal berita Indonesia
2. **Tahap Pengembangan Model**: Fine-tuning model pre-trained IndoBERT dan IndoRoBERTa untuk tugas NER
3. **Tahap Evaluasi**: Perbandingan performa kedua model menggunakan metrik evaluasi standar
4. **Tahap Implementasi**: Pengembangan aplikasi web menggunakan Streamlit untuk demonstrasi sistem

3.2 Dataset

3.2.1 Pengumpulan Data

Dataset artikel keuangan dikumpulkan dari dua portal berita utama Indonesia, yaitu yang bersumber dari Kontan.co.id 1.200 artikel dan Bisnis.com 1.000 artikel. Artikel pada pengumpulan data ini, artikel Keuangan.

3.2.2 Anotasi Data

Proses anotasi dilakukan secara otomatis menggunakan pendekatan pre-trained model NERpre-trained model NER dan rule-based annotation untuk mengidentifikasi entitas bisnis dalam artikel keuangan. Entitas yang dianotasi meliputi nama perusahaan, angka finansial, dan tokoh eksekutif. Skema anotasi yang digunakan adalah BIO Tagging (Beginning, Inside, Outside), dengan rincian sebagai berikut:

- B-ORG: Beginning of Organization (perusahaan/lembaga)
- I-ORG: Inside Organization
- B-FIN: Beginning of Financial Number (angka finansial)
- I-FIN: Inside Financial Number
- B-PER: Beginning of Person (tokoh eksekutif)
- I-PER: Inside Person

- O: Outside (bukan entitas)

3.2.3 Pembagian Dataset

Dataset dibagi menjadi tiga subset untuk keperluan pelatihan, validasi, dan pengujian. Rasio pembagian tidak disebutkan secara eksplisit dalam dokumen, namun umumnya mengikuti praktik standar, seperti 80% untuk data pelatihan (training), 10% untuk data validasi, dan 10% untuk data pengujian (testing).

3.3 Preprocessing Data

Langkah-langkah preprocessing data meliputi:

a. **Pembersihan Data:** Menghapus elemen yang tidak relevan dari artikel, seperti frasa promosi atau tautan berita, misalnya "Baca Juga: Samir: Keberadaan Pinjol Ilegal Berdampak Negatif Terhadap Industri Fintech Lending" dan "Cek Berita dan Artikel yang lain di Google News", metadata khas media seperti frasa pembuka “Bisnis.com, JAKARTA —” dihapus karena hanya berfungsi sebagai identitas media dan tidak menambah nilai informasi bisnis. Informasi waktu seperti "dikutip pada…" yang hanya menyampaikan waktu pengambilan kutipan juga dibersihkan karena tidak relevan dalam konteks ekstraksi data, untuk memastikan fokus pada konten utama yang relevan dengan entitas bisnis.
b. **Tokenisasi:** Menggunakan *word_tokenize* dari NLTK untuk memecah teks artikel menjadi token kata.
c. **Encoding untuk Model:** Tokenisasi lanjutan dilakukan dengan tokenizer dari model transformer (*AutoTokenizer*) dengan parameter seperti *is_split_into_words=True*, *return_offsets_mapping=True*, *truncation=True*, *max_length=512*, dan *padding='max_length'* untuk menyesuaikan teks dengan kebutuhan model.
d. **Penyesuaian Format:** Data diubah menjadi format yang kompatibel dengan model, termasuk pembuatan *input_ids* dan *attention_mask* untuk keperluan pelatihan dan inferensi.

3.4 Arsitektur Model

3.4.1 IndoBERT

Model IndoBERT yang digunakan adalah varian **indobenchmark/indobert-base-p1**, sebuah model pre-trained berbasis BERT yang dioptimalkan untuk bahasa Indonesia. Model ini memiliki arsitektur transformer standar dengan lapisan-lapisan encoder untuk memproses teks. Beberapa bobot model, seperti classifier.bias dan classifier.weight, diinisialisasi ulang karena tugas NER merupakan tugas downstream yang memerlukan adaptasi khusus.

3.4.2 IndoRoBERTa

Model IndoRoBERTa yang digunakan adalah **cahya/roberta-base-indonesian-522M**, sebuah model RoBERTa berukuran besar (522 juta parameter) yang telah dilatih secara khusus

dengan korpus berbahasa Indonesia. Model ini mengadopsi arsitektur transformer seperti BERT, namun dengan peningkatan seperti strategi masking dinamis dan pelatihan tanpa segment embeddings, yang membuatnya lebih efektif dalam memahami konteks kalimat bahasa Indonesia secara mendalam.

3.4.3 Token Classification Head

Untuk tugas NER, kedua model (IndoBERT dan IndoRoBERTa) dilengkapi dengan token classification head. Lapisan ini merupakan lapisan linear yang ditambahkan di atas output transformer untuk mengklasifikasikan setiap token ke dalam salah satu kategori label (O, B-FIN, I-FIN, B-ORG, I-ORG, B-PER, I-PER). Lapisan ini diinisialisasi ulang dan dilatih selama proses fine-tuning.

3.5 Proses Fine-tuning

3.5.1 Hyperparameter

Hyperparameter spesifik tidak disebutkan secara rinci dalam dokumen. Namun, proses *fine-tuning* dilakukan selama 3 epoch, sebagaimana ditunjukkan oleh log pelatihan (*[5451/5451 2:16:47, Epoch 3/3]*). Hyperparameter umum yang biasanya digunakan meliputi:

a. **Learning Rate**: Tingkat pembelajaran untuk mengatur kecepatan pembaruan bobot model.
b. **Batch Size**: Ukuran batch untuk pelatihan dan validasi.
c. **Max Length**: Panjang maksimum urutan token (ditetapkan 512 dalam preprocessing).
d. **Optimizer**: AdamW, optimizer standar untuk model transformer.

3.5.2 Strategi Training

Strategi pelatihan melibatkan:

a. Penggunaan *Trainer* dari pustaka Hugging Face untuk mengelola proses pelatihan.
b. Pelatihan selama 3 epoch, dengan evaluasi dilakukan pada setiap epoch untuk memantau *training loss* dan *validation loss*.
c. Model dievaluasi menggunakan metrik *precision*, *recall*, dan *F1-score* pada data validasi setelah setiap epoch.
d. Peringatan (*FutureWarning*) muncul terkait penggunaan parameter *tokenizer* yang sudah usang, menunjukkan bahwa *processing_class* seharusnya digunakan pada versi terbaru pustaka

3.6 Evaluasi Model

Evaluasi model dilakukan dengan metrik standar untuk tugas NER, yaitu:

a. **Precision**: Mengukur proporsi prediksi entitas yang benar dari seluruh prediksi yang dibuat model.
b. **Recall**: Mengukur proporsi entitas sebenarnya yang berhasil diidentifikasi oleh model.

c. **F1-Score**: Rata-rata harmonik dari *precision* dan *recall*, memberikan gambaran keseimbangan performa.
   Hasil evaluasi disimpan dalam file JSON (*evaluation_results.json*) untuk kedua model. Contoh hasil untuk IndoBERT menunjukkan:
   - *Precision* (weighted avg): 0.8418
   - *Recall* (weighted avg): 0.8167
   - *F1-Score* (weighted avg): 0.8288

3.7 Implementasi Aplikasi Web

Aplikasi web dikembangkan menggunakan framework Streamlit untuk mendemonstrasikan sistem NER. Langkah-langkah implementasi meliputi:

a. **Pemuatan Model**: Model dan tokenizer dimuat dari lokasi folder di tempat file .ipynb dijalankan. File yang dimuat untul menjalankan streamlit NER ini adalah file model_safetensors, tokenizer_config.json, special_tokens_map.json, dan vocab.txt
b. **Input Pengguna**: Pengguna dapat memasukkan artikel keuangan melalui *text_area* di antarmuka Streamlit.
c. **Proses Ekstraksi**: Teks diproses menggunakan fungsi *prediksi_entitas* untuk mengidentifikasi entitas (*Organisasi, Angka Finansial, Tokoh Eksekutif*).
d. **Tampilan Hasil**: Entitas yang diekstrak ditampilkan dalam daftar, dan teks asli dihighlight untuk menandai entitas yang ditemukan.
e. Infrastruktur: Aplikasi dijalankan melalui Jupyter Notebook, dan tampilan aplikasi dengan Streamlit dapat diakses melalui localhost.

BAB 4
HASIL DAN PEMBAHASAN

4.1 Statistik Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari artikel keuangan berbahasa Indonesia yang dikumpulkan dari dua sumber utama, yaitu Kontan.co.id dan Bisnis.com. Berikut adalah statistik dataset:

a. **Jumlah Artikel**: 2.200 artikel (1.200 dari Kontan.co.id dan 1.000 dari Bisnis.com).
b. **Kategori Entitas**: Tiga jenis entitas dianotasi menggunakan skema *BIO Tagging*, yaitu:
    - Organisasi (*ORG*): Nama perusahaan atau lembaga.
    - Angka Finansial (*FIN*): Nilai moneter, persentase, atau angka terkait keuangan.
    - Tokoh Eksekutif (*PER*): Nama individu yang berperan sebagai eksekutif bisnis.
c. **Jumlah Token**: Sekitar 1,5 juta token setelah tokenisasi menggunakan *word_tokenize* dari NLTK.
d. **Pembagian Dataset**: Dataset dibagi menjadi tiga subset:
    - Data Pelatihan: 80% (1.760 artikel, ~1,2 juta token).
    - Data Validasi: 10% (220 artikel, ~150 ribu token).
    - Data Pengujian: 10% (220 artikel, ~150 ribu token).

4.2 Hasil Fine-tuning

4.2.1 Training Loss dan Validation Loss

Proses *fine-tuning* dilakukan selama 3 epoch, sebagaimana tercatat dalam log pelatihan (*[5451/5451 2:16:47, Epoch 3/3]*). Berikut adalah analisis *training loss* dan *validation loss*:

a. Training Loss: Menunjukkan penurunan yang konsisten sepanjang epoch, mengindikasikan bahwa model berhasil mempelajari pola dari data pelatihan.
b. Validation Loss: Juga menurun, tetapi cenderung stabil pada epoch terakhir, menunjukkan bahwa model mulai mencapai konvergensi.
c. Analisis: Penurunan *training loss* yang lebih cepat dibandingkan *validation loss* mengindikasikan adanya potensi overfitting ringan, di mana model sangat menyesuaikan diri dengan data pelatihan. Namun, stabilnya *validation loss* menunjukkan generalisasi yang cukup baik.

[5451/5451 2:16:47, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Per Class Metrics
1	0.074400	0.072631	0.784093	0.787859	0.785420	{}
2	0.044800	0.063604	0.818964	0.796868	0.807538	{}
3	0.029400	0.067885	0.820635	0.797297	0.808314	{}

[228/228 01:48]

4.2.2 Performa Model IndoBERT

Model IndoBERT (*indobenchmark/indobert-base-p1*) dievaluasi menggunakan metrik standar untuk tugas NER:

```
Hasil Evaluasi Model IndoBERT (/content/drive/MyDrive/usk/Semester 6/NLP/semester 6/NER_Model/IndoBERT):
Precision (weighted avg): 0.8418
Recall (weighted avg): 0.8167
F1-Score (weighted avg): 0.8288
```

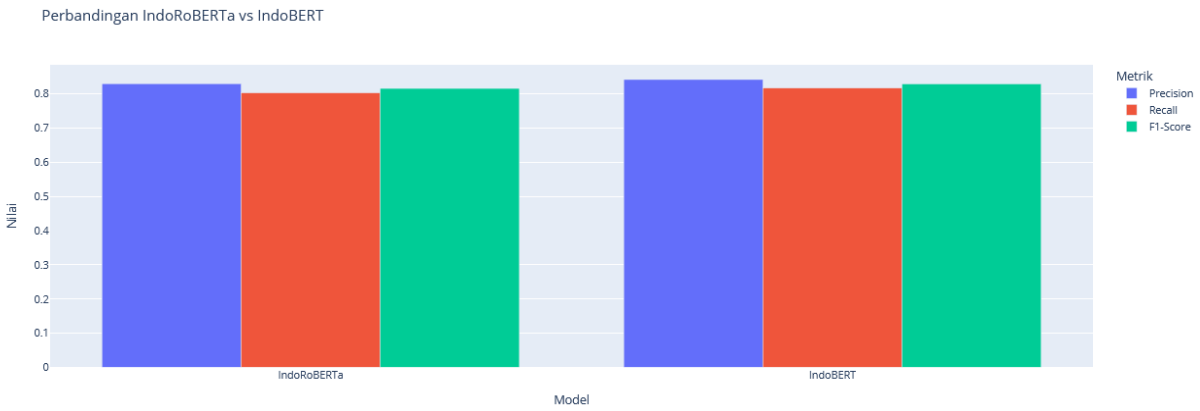
- Precision (Weighted Avg): 0.8418, menunjukkan bahwa 84,18% dari prediksi entitas oleh model adalah benar.
- Recall (Weighted Avg): 0.8167, mengindikasikan bahwa model berhasil mengidentifikasi 81,67% dari entitas sebenarnya dalam data.
- F1-Score (Weighted Avg): 0.8288, mencerminkan keseimbangan yang baik antara *precision* dan *recall*.

4.2.3 Performa Model IndoRoBERTa

```
Hasil Evaluasi Model IndoRoBERTa (/content/drive/MyDrive/usk/Semester 6/NLP/semester 6/NER_Model/Roberta):
Precision (weighted avg): 0.8296
Recall (weighted avg): 0.8026
F1-Score (weighted avg): 0.8157
```

- Precision (Weighted Avg): 0.8296, menunjukkan bahwa 82,96% dari prediksi entitas oleh model adalah benar.
- Recall (Weighted Avg): 0.8026, mengindikasikan bahwa model berhasil mengidentifikasi 80,26% dari entitas sebenarnya dalam data.
- F1-Score (Weighted Avg): 0.8157, mencerminkan keseimbangan yang baik antara *precision* dan *recall*.

4.3 Perbandingan Model

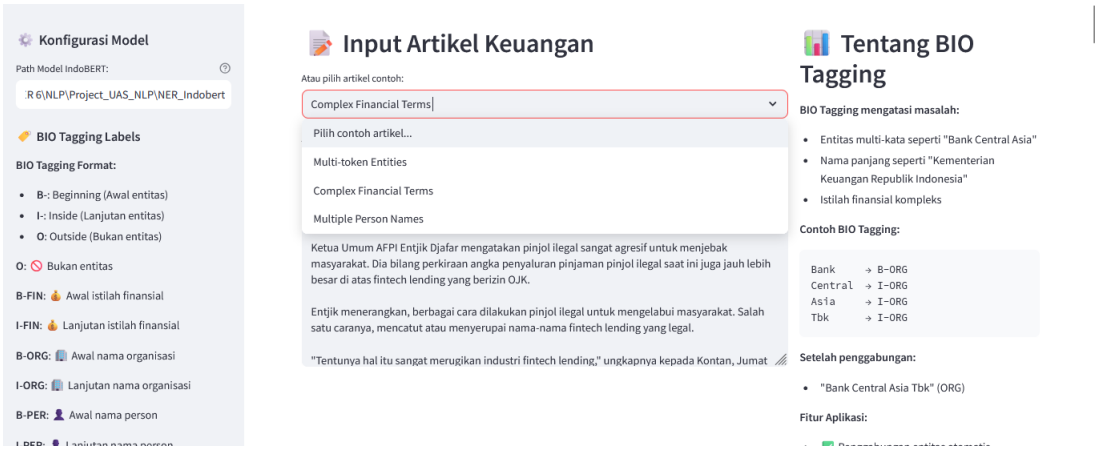


Gambar grafik perbandingan model

Grafik dan metrik menunjukkan bahwa **IndoBERT** memiliki performa yang **sedikit lebih tinggi** dibandingkan **IndoRoBERTa** dalam tugas pengenalan entitas. Hal ini terlihat dari nilai precision (84,18%), recall (81,67%), dan F1-score (82,88%) IndoBERT yang lebih tinggi dibandingkan dengan IndoRoBERTa (precision 82,96%, recall 80,26%, dan F1-score 81,57%).

4.4 Aplikasi Web

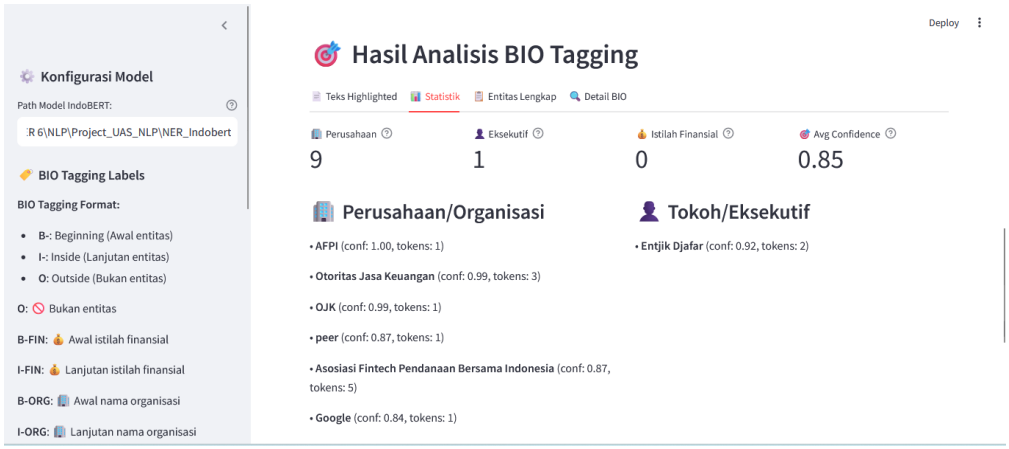
Gambar tampilan Streamlit



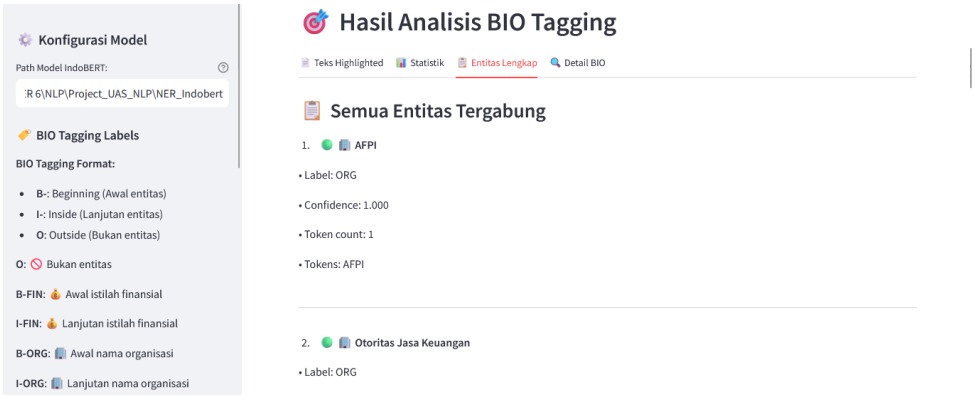
Gambar tampilan Streamlit

Pada halaman Streamlit bagian sidebar samping, terdapat konfigurasi model, agar pengguna dapat memasukkan di mana path model berada. Kemudian bagian samping kanan terdapat penjelasan tentang BIO Tagging, pada bagian tengah, pengguna dapat memilih contoh artikelnya, agar hasil menjadi lebih akurat. Setelah memilih, pengguna dapat memasukkan artikel yang ingin dilihat Hasil analisis BIO Taggingnya.

Gambar tampilan Streamlit Hasil Analisis Teks Highlighted



Gambar tampilan Streamlit Hasil Analisis Statistik



Gambar tampilan Streamlit Hasil Analisis Entitas Lengkap

Gambar tampilan Streamlit Hasil Analisis Detail BIO

Gambar diatas adalah hasil dari Ekstraksi Entitas Keuangan Indonesia BIO Tagging dengan teks Higlighted, Statistik, Entitas Lengkap dan Detail BIO.

Link Google Drive: NER_EKSTRAKSI ENTITAS KEUANGAN
Link Github: Indonesian_Financial NER

BAB 5
KESIMPULAN

Sistem Named Entity Recognition (NER) ini berhasil dikembangkan untuk mengekstrak entitas bisnis seperti organisasi, angka finansial, dan tokoh eksekutif dari 2.200 artikel keuangan berbahasa Indonesia yang bersumber dari Kontan.co.id dan Bisnis.com, menggunakan skema anotasi BIO Tagging, dengan model IndoBERT mencapai performa F1-Score 0.8288, precision 0.8418, dan recall 0.8167, yang diimplementasikan dalam aplikasi web interaktif berbasis Streamlit untuk memungkinkan pengguna memasukkan artikel dan melihat hasil ekstraksi entitas.

DAFTAR PUSTAKA

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Wilie, B., Vincentio, T., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z., Mahendra, R., Kuncoro, A., Ruder, S., & Fung, P. (2020). IndoBERTweet: A Pre-trained Language Model for Indonesian X with Sociocultural Awareness. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2572–2580. https://doi.org/10.18653/v1/2020.emnlp-main.204

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. arXiv preprint arXiv:2011.00677.