1. Pendahuluan

Latar Belakang

Sebagai kitab suci umat Islam, Alquran diturunkan menggunakan Bahasa Arab yang memiliki struktur bahasa yang kompleks, termasuk sistem penulisan yang berbeda, tata bahasa yang rumit, dan sistem fonetik yang berbeda dari bahasa-bahasa lain yang umumnya dipelajari[1]. Teks Alquran terdiri lebih dari 6200 ayat dan 114 surah[2]. Selain itu, gaya bahasanya yang kaya dan maknanya yang mendalam[3] membuatnya sulit dipahami. Pemahaman dan interpretasi terhadap ayat-ayat Alquran merupakan unsur tafsir dari terjemahan Alquran[4] yang menjadi tujuan utama bagi umat Islam. Umat Islam membutuhkan pemahaman tentang entitas manusia dalam Alquran untuk memahami ajaran yang terkandung di dalamnya. Sangat penting untuk menjelaskan pesan dan nilai-nilai yang terkandung dalam setiap ayat dengan menggunakan entitas manusia. Namun, proses pelabelan entitas manusia pada Alquran secara manual dapat menjadi tugas yang rumit dan dapat menyebabkan kesalahan. Pendekatan yang lebih efektif diperlukan untuk mengidentifikasi dan melabeli objek karena ukuran teks yang besar dan kompleksitas bahasanya.

Salah satu cara yang dapat dilakukan untuk memahami dan mengidentifikasi entitas manusia adalah Ekstraksi Informasi. Ekstraksi Informasi (IE) adalah salah satu tipe pemrosesan dokumen yang menangkap dan menghasilkan informasi faktual yang tersedia di dalam dokumen tersebut[5]. Salah satu masalah dalam IE adalah informasi yang diolah umumnya tidak terstruktur. *Named Entity Recognition* (NER) adalah salah satu pendekatan yang digunakan untuk menangani masalah ini. NER merupakan proses mengekstraksi entitas bernama yang dianggap penting dalam sebuah teks yang menentukan kategorinya ke dalam kategori yang telah terdefinisi[6]. Dalam Alquran, NER dapat digunakan untuk mengidentifikasi dan melabeli entitas manusia seperti nama karakter atau kelompok.

Beberapa tahun terakhir, model NER berbasis *deep learning* menjadi dominan dan mencapai hasil yang canggih[7]. *Deep learning* adalah bidang *machine learning* yang terdiri dari beberapa lapis pemrosesan untuk mempelajari representasi data dengan berbagai tingkat abstraksi. Keuntungan utama *deep learning* adalah kemampuan pembelajaran represetasi dan komposisi semantik yang didukung oleh representasi vektor dan pemrosesan saraf[7]. Ini memungkinkan mesin untuk menerima data mentah dan secara otomatis menemukan pemrosesan dan representasi laten yang diperlukan untuk klasifikasi atau deteksi.

BERT (*Bidirectional Encoder Representation from Transformers*), yaitu model representasi bahasa baru yang dirancang untuk melatih representasi dua arah yang mendalam dari teks yang tidak berlabel dengan melakukan pengecekan bersama pada konteks kiri dan kanan di semua lapisan[8]. Pada penelitian sebelumnya oleh Retno Diah, dkk., telah dilakukan ekstraksi entitas manusia pada Alquran terjemahan Bahasa Inggris menggunakan BERT. Model dibangun dengan menggunakan dataset dari Quran Surah Al-Fatihah, Al-Baqarah, dan Ali Imran. Penelitian ini menghasilkan performa *F1-Score* sebesar 53% dengan penggunaan *hyperparameter learning rate* 0,0001, *epoch* 85, dan *batch size* 32. Hasil menunjukkan bahwa model yang dibangun memiliki kesulitan dalam mengklasifikasikan frasa bertingkat karena kurangnya data latih sehiingga menghasilkan performa yang rendah untuk melakukan ekstraksi manusia[8].

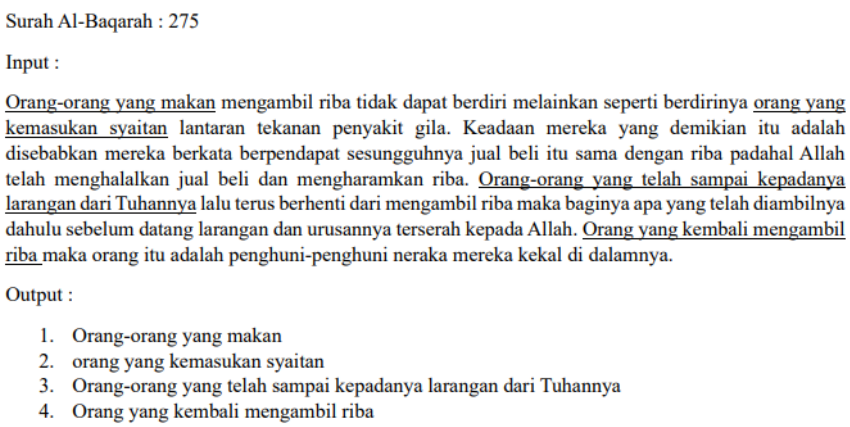
RoBERTa (*A Robustly Optimized BERT Pretraining Approach*) adalah model pengembangan BERT yang dirilis pada tahun 2019 oleh peneliti dari University of Washington dan Facebook AI[9]. RoBERTa adalah BERT yang dilatih ulang dengan metodologi pelatihan yang ditingkatkan. Hasil penelitian yang dilakukan menunjukkan bahwa, saat dilakukan pra-pelatihan dengan *dataset* yang sama dengan BERT, RoBERTa lebih akurat 1000% pada dataset SQUAD, MNLI-m, dan SST-2 dan membutuhkan langkah sepuluh kali lebih sedikit daripada BERT. Sedangkan, penggunaan RoBERTa untuk mengekstraksi entitas manusia pada terjemahan Alquran Bahasa Indonesia belum pernah dilakukan.

Penelitian ini bertujuan memudahkan proses pelabelan entitas manusia dalam teks Alquran dengan menggunakan model RoBERTa. RoBERTa memiliki kemampuan pemrosesan bahasa yang lebih baik karena dilatih pada data yang lebih besar dan lebih beragam dari model sebelumnya seperti BERT dan XLNet[9]. RoBERTa juga memiliki performa yang lebih baik pada tugas NER dibandingkan dengan model lain seperti BERT dan XLNet[9]. Dengan demikian RoBERTa dapat membangun model dengan hasil yang lebih akurat dalam pelabelan entitas manusia dalam Alquran Bahasa Indonesia.

Dalam pembuatan model RoBERTa digunakan dataset dari korpus Tanzil Quran berupa teks terjemahan Alquran Bahasa Indonesia. Penelitian ini melabeli entitas manusia menggunakan format BIO, yaitu *Begin* (B) menandakan awal dari satu frase label, *Inside* (I) menandakan bagian dari suatu frase label, dan *Outside* (O) menandakan tidak termasuk frase label yang ditentukan[10]. Tabel 1 menunjukkan struktur dataset yang digunakan dalam penelitian ini. Input dari sistem ekstraksi entitas manusia yang dibangun adalah kalimat atau teks terjemahan ayat Alquran Bahasa Indonesia dan output dari sistem adalah entitas manusia yang terekstraksi oleh sistem. Gambar 1 menunjukkan contoh input yang diambil dari potongan ayat Alquran Surat Al-Baqarah ayat 275.

Tabel 1. Struktur dataset

Juz	Ayat	Token	Label
1	7	yaitu	O
1	7	jalan	O
1	7	orang-orang	B-PER
1	7	yang	I-PER
1	7	telah	I-PER
1	7	engkau	I-PER
1	7	beri	I-PER
1	7	nikmat	I-PER



Surah Al-Baqarah : 275

Input :

Orang-orang yang makan mengambil riba tidak dapat berdiri melainkan seperti berdirinya orang yang kemasukan syaitan lantaran tekanan penyakit gila. Keadaan mereka yang demikian itu adalah disebabkan mereka berkata berpendapat sesungguhnya jual beli itu sama dengan riba padahal Allah telah menghalalkan jual beli dan mengharamkan riba. Orang-orang yang telah sampai kepadanya larangan dari Tuhannya lalu terus berhenti dari mengambil riba maka baginya apa yang telah diambilnya dahulu sebelum datang larangan dan urusannya terserah kepada Allah. Orang yang kembali mengambil riba maka orang itu adalah penghuni-penghuni neraka mereka kekal di dalamnya.

Output :

1. Orang-orang yang makan
2. orang yang kemasukan syaitan
3. Orang-orang yang telah sampai kepadanya larangan dari Tuhannya
4. Orang yang kembali mengambil riba

Gambar 1. Contoh input dan output sistem ekstraksi entitas manusia

Topik dan Batasannya

Penelitian ini difokuskan pada pelabelan entitas manusia pada teks terjemahan Alquran Bahasa Indonesia dengan tujuan meningkatkan pemahaman tentang entitas manusia dalam Alquran. Dalam penerapannya, proses pelabelan seringkali melibatkan tantangan yang kompleks. Mengidentifikasi dan mengkategorikan entitas manusia seperti nama karakter atau kelompok dalam teks memerlukan kecermatan untuk memahami konteks ayat-ayat Alquran.

Penelitian ini membatasi dataset yang digunakan untuk melatih model dengan hanya menggunakan terjemahan Alquran juz 1 sampai dengan 6 yang sudah terlabeli sebelumnya. Entitas yang dilabeli dalam penelitian ini terkhusus pada entitas manusia saja. Batasan lainnya terkait model yang digunakan sebagai perbandingan adalah model BERT.

Tujuan

Penelitian ini dilakukan dengan membangun model RoBERTa untuk mengidentifikasi entitas manusia pada teks terjemahan Alquran Bahasa Indonesia. Penelitian ini juga bertujuan untuk mengetahui perbandingan performa model RoBERTa dengan model lain dalam mengidentifikasi entitas manusia pada teks terjemahan Alquran Bahasa Indonesia.

Organisasi Tulisan

Pada bab 2, dijelaskan studi literatur yang mencakup teori, metrik pengukuran, dan data yang digukanan pada penelitian ini. Pada bab 3, dipaparkan mengenai perancangan sistem dengan menggunakan model RoBERTa. Pada bab 4 berisi evaluasi hasil pengujian yang dilakukan beserta analisis dari hasil tersebut. Dan, pada bab 5 berisi hasil kesimpulan dari penelitian yang dilakukan beserta saran dan rekomendasi untuk penelitian selanjutnya.