BAB I
PENDAHULUAN

1.1. Latar Belakang

Saat ini, di era Industri 4.0, *Artificial Intelligence* (AI) dimanfaatkan di banyak sektor dan kasus studi untuk membuat aktivitas manusia lebih mudah dan lebih efisien. AI adalah teknologi yang membuat komputer lebih cerdas, tidak hanya untuk komputasi, tetapi juga untuk memprediksi, mendeteksi, mengenali, menganalisis, dan melakukan aktivitas seperti yang dilakukan manusia. Sehingga AI mampu dan handal untuk menyelesaikan banyak kasus di bidang kesehatan, ekonomi, dan bisnis yang biasa disebut dengan bisnis cerdas, *game*, pendidikan, robotika, dan AI umum teknologi dipadukan dengan internet yang populer dengan istilah *Internet of Things* (IoT). Selain itu, untuk pemrosesan bahasa, ada teknik khusus dalam AI yang disebut *Natural Language Processing* (NLP). NLP adalah proses untuk menganalisis dan menemukan wawasan pengetahuan data yang mengandung bahasa, seperti data teks dan data ucapan NLP adalah bidang AI yang mempelajari komunikasi antara komputer dan manusia melalui bahasa alami [1].

Jumlah informasi di web telah tumbuh secara eksponensial selama bertahun-tahun, dengan konten yang mencakup hampir semua topik. Akibatnya, ketika mencari informasi, pengguna sering bingung dengan banyaknya hasil dari mesin pencari. Pengguna biasanya harus telaten menelusuri daftar panjang hasil untuk mencari jawaban yang tepat. Oleh karena itu penelitian *Question Answering* (QA) muncul dalam upaya untuk mengatasi masalah informasi yang berlebihan ini. Alih-alih mengembalikan daftar peringkat hasil seperti yang dilakukan di mesin pencari saat ini, QA bertujuan untuk memanfaatkan analisis konten linguistik dan media yang mendalam serta pengetahuan domain untuk mengembalikan jawaban yang tepat untuk pertanyaan bahasa alami [2].

Question Answering adalah sistem yang memungkinkan pengguna untuk mengajukan pertanyaan dengan cara yang alami dan spesifik. Sistem ini akan mengembalikan daftar dokumen teks singkat atau frasa sebagai jawaban, yang kemudian dapat disaring oleh pengguna untuk menentukan apakah dokumen tersebut berisi jawaban yang sesuai. Dengan menggunakan sistem *Question*

Answering, pengguna dapat menanyakan pertanyaan dalam bahasa sehari-hari dan menerima jawaban yang cepat, singkat, dan dapat diperkuat dengan kalimat yang mendukung kebenaran dari jawaban tersebut. Secara umum, sistem *Question Answering* terdiri dari enam tahapan proses, yaitu analisis pertanyaan, pra-pemrosesan dokumen, pemilihan dokumen kandidat, analisis dokumen kandidat, ekstraksi jawaban, dan pembuatan respons [3]. Salah satu model yang dapat digunakan untuk menerapkan *Question Answering* adalah model XLM-RoBERTa.

XLM-RoBERTa merupakan sebuah *pre-trained model* penyandi kalimat multibahasa berskala, *pre-trained model* ini dilatih dengan 2,5 Terabyte data dalam 100 bahasa data yang difilter dari *Common Crawl*. XLM-RoBERTa mencapai hasil terbaik pada berbagai tolok ukur multibahasa. Model XLM-RoBERTa diusulkan dalam Pembelajaran Representasi Multibahasa Tanpa Pengawasan atau *Unsupervised Learning* pada Skala oleh Alexis Conneau, dkk. Model ini menggunakan teknik *mask language modeling* dalam melakukan pelatihan kemudian model ini juga dapat dilatih menjadi spesifik *task* dari *natural language processing* seperti *text classification, question answering, named entity recognition, text generation,* dan lain-lain menggunakan dataset sesuai kebutuhan dari masing-masing spesifik task *natural language processing* [4]. Model ini dibuat berdasarkan pada model RoBERTa Facebook yang dirilis pada 2019, model ini merupakan model bahasa multibahasa yang besar. Model XLM-RoBERTa ini juga mengungguli mBERT pada klasifikasi lintas bahasa dengan akurasi hingga melebihi 23% pada bahasa sumber daya rendah yang mengungguli keadaan sebelumnya dengan akurasi rata-rata 5,1% pada XNLI, 2,42% rata-rata F1-score on *Named Entity Recognition*, dan data rata-rata 9,1%, dibandingkan dengan metode yang digunakan. F1-skor pada menjawab pertanyaan lintas bahasa. Model ini juga mengevaluasi *fine tuning* monolingual pada tolok ukur GLUE dan XNLI, di mana XLM-R memperoleh hasil yang bersaing dengan model monolingual canggih, termasuk RoBERTa [5]. Hasil ini menunjukkan, untuk pertama kalinya, bahwa adalah mungkin untuk memiliki satu model besar untuk semua bahasa.

Penelitian sebelumnya yang dilakukan oleh Leonardus (2015) telah dibangun *question answering system* untuk pertanyaan *factoid* dengan studi kasus

biografi Presiden Indonesia ke-1 hingga ke-7 dan Wakil Presiden Indonesia ke-1 hingga ke-12 menggunakan Wordnet, untuk melakukan proses ekstraksi jawaban Selain itu juga digunakan Dice Coefficient untuk menghitung kemiripan dokumen dengan query untuk meningkatkan efektivitas *question answering system* dengan hasil menunjukkan bahwa *question answering system* setelah menggunakan WordNet (dengan 0.7083 mrr dan 0.6577 mrr dan 0.6577 mrr) tidak dapat meningkatkan performa sistem sebelum menggunakan WordNet dengan 0.7083 mrr [6].

Penelitian ini akan mengimplementasikan dataset bahan ajar berupa dataset validasi SQuAD 2.0, dan dataset sejarah UIN. Sistem yang akan dibangun kali ini yaitu dikhususkan untuk pertanyaan *factoid* (apa, siapa, kapan, dimana) dengan kembalian jawaban berupa suatu kalimat atau kata. Namun kali ini akan menggunakan sebuah *pre – trained model* terhadap *question answering system* dengan objek yang digunakan yaitu dataset bahan ajar dan akan menggunakan *pre – trained model Cross-lingual Language Model Robustly Optimized BERT Pre-training Approach* (XLM-RoBERTa).

Pada pemaparan latar belakang diatas untuk mengetahui apakah *pre – trained model* XLM – RoBERTa dapat melakukan *Question Answering* pada dataset bahan ajar maka perlu dilakukan penelitian. Dengan permasalahan tersebut maka dirumuskan penelitian yang berjudul "**Implementasi Model XLM-RoBERTa Pada *Question Answering* Dataset SQuAD 2.0**", penelitian ini diharapkan menghasilkan model yang dapat melakukan *question answering* terhadap dataset bahan ajar dengan akurat yang nantinya akan ditampilkan dalam sebuah aplikasi pada platform web.

1.2. Rumusan Masalah

1. Bagaimana menerapkan model XLM-RoBERTa pada pembangunan aplikasi *question answering* pada dataset bahan ajar?.
2. Bagaimana kinerja model XLM-RoBERTa pada pembangunan aplikasi *question answering* pada dataset bahan ajar?.

1.3. Tujuan dan Manfaat Penelitian

1 Menerapkan model XLM-RoBERTa pada pembangunan aplikasi *question answering* pada dataset bahan ajar.

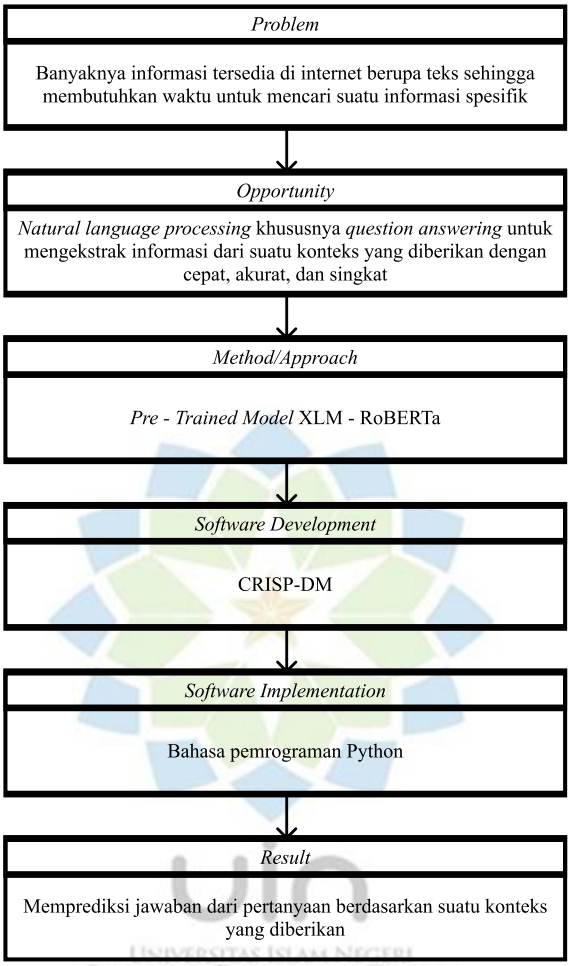
2 Mengetahui kinerja model XLM-RoBERTa pada pembangunan aplikasi *question answering* pada dataset bahan ajar.

1.4. Batasan Masalah

1. Sistem hanya menerima masukan berupa pertanyaan dalam bahasa Indonesia dalam bentuk *factoid* (apa, dimana, berapa, dan kapan).
2. Sistem akan dibangun 3 model berbeda menggunakan dataset SQuAD 2.0 atau *Stanford Question Answering Dataset* yang tersedia dan telah diterjemahkan ke dalam bahasa Indonesia dengan versi ukuran kecil dan besar.
3. Sistem akan mengimplementasikan dataset bahan ajar berupa dataset validasi SQuAD 2.0, dan dataset sejarah UIN.
4. Sistem akan mengembalikan jawaban berupa kalimat atau kata yang berasal dari konteks diberikan.
5. Sistem akan menggunakan *extractive question answering* daripada *generative question answering*.
6. Sistem akan ditampilkan dalam bentuk aplikasi pada platform web.
7. Sistem akan dibangun menggunakan bahasa python.
8. Sistem akan diukur menggunakan *confusion matrix*.

1.5. Kerangka Pemikiran

Kerangka pemikiran pada penelitian ini ditunjukkan pada gambar 1.1



Gambar 1.1 Kerangka Pemikiran

1.6. Sistematika Penulisan

Sistematika penulisan pada penelitian kali ini dibagi ke dalam 5 bab. Pada setiap bab dijelaskan sesuai tujuan dan pengembangan dari sistem tersendiri. Sistematika penulisan pada penelitian ini dapat dilihat sebagai berikut:

BAB I : Pendahuluan

Bab I berisikan tentang latar belakang penelitian, rumusan masalah, tujuan, batasan, metode pengembangan sistem, kerangka hingga kerangka pemikiran. Berikut dengan sistematika penulisan disajikan.

BAB II : Kajian Literature

Bab II menjelaskan tentang pembahasan penelitian terdahulu serta konsep-konsep dan teori pendukung pada penelitian yang akan dilakukan.

BAB III : Metodologi Penelitian

Bab III berisikan tentang metode yang digunakan dalam penyusunan tugas akhir. Metodologi penelitian disajikan berdasarkan analisis kebutuhan menggunakan metode CRISP-DM. Dalam metode tersebut beberapa tahapan yagn terdapat pada Bab III ini adalah Pemahaman Bisnis, Pemahaman Data, Persiapan Data dan *Modeling Phase.*

BAB IV : Hasil dan Pembahasan

Bab ini membahas mengenai hasil dari implementasi sistem itu sendiri seperti hasil dari perhitungan training dan pengujian yang dilakukan, bab ini berisi lanjutan tahapan CRISP-DM dari bab sebelumnya yaitu *Evaluation Phase.*

BAB V : Simpulan dan Saran

Bab ini berisi kesimpulan dari penelitian yang dilakukan serta saran yang direkomendasikan untuk peningkatan atau perbaikan dari penelitian ini.