BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dari penelusuran informasi secara daring hingga pencarian informasi, kegiatan menjawab pertanyaan menjadi lumrah dan diterapkan secara luas dalam kehidupan sehari-hari. Dalam tugas *Question Answering*, dengan mempertimbangkan perkembangan teknologi komputer saat ini, mesin cenderung dapat memproses informasi dalam jumlah yang lebih banyak sehingga dapat menjawab pertanyaan lebih cepat terhadap suatu domain tertentu [1]. Hal demikian memungkinkan mesin untuk dapat menjadi kandidat yang tepat dalam mendapatkan konteks dari suatu bacaan dengan tujuan untuk memberikan jawaban yang paling akurat terhadap konteks pertanyaan yang diajukan [2].

Pemanfaatan model bahasa berbasis *deep learning* seperti BERT mengizinkan mesin untuk dapat memahami konteks suatu kalimat bahasa alami (*natural language*) sehari-hari, sehingga dengan kapabilitas tersebut membuka peluang dalam otomasi beragam tugas-tugas yang dapat diselesaikan secara NLP (*Natural Language Processing*) seperti tanya-jawab, mendapatkan kesimpulan, parafrasa dan lain-lainnya. Pada studi model bahasa BERT (*Bidirectional Encoder Representations from Transformers*) dikaji bahwa suatu model bahasa yang melakukan *training* secara *bidirectional* (*left-to-right* dan *right-to-left*) dapat lebih baik dalam menangkap konteks pada data dengan kalimat yang lebih panjang dibandingkan dengan model bahasa satu arah. Teknik ini dikenal dengan *Masked Language Model* yang mengizinkan *Bidirectional Training* [3]. Di samping itu, BERT juga berbasis arsitektur *attention* (*transformer*) [4] yang mampu menghasilkan hasil *state-of-the-art* (SOTA) terhadap berbagai *task Natural Language Processing* (NLP) dibandingkan dengan model berbasis arsitektur *Recurrent Neural Network* (RNN) seperti *Bidirectional Long Short Term Memory* (BiLSTM) dan *Bidirectional Gated Recurrent Unit* (BiGRU).

Kemudian terdapat beberapa penelitian model bahasa dengan mekanisme *attention* (*transformer*) lainnya seperti model bahasa T-NLG (Turing *Natural Language Generation*) yang dilatih dengan 17 miliar parameter [5] dan model bahasa GPT-3 (*Generative Pre-trained Transformer* 3) yang di*train* dengan 175 miliar parameter [6] sehingga dapat menghasilkan *state-of-the-art* (SOTA) terhadap berbagai

benchmark pemodelan bahasa serta berbagai *task* seperti *summarization* dan *question answering*.

Model bahasa yang berbasis arsitektur *transformer* seperti BERT, dan GPT umumnya mampu mencapai hasil yang *state-of-the-art* dalam berbagai *task* NLP. Namun model bahasa seperti BERT yang menggantikan token *random* dengan *mask* serta di-*encode* secara dua arah hanya terfokus untuk domain *Natural Language Understanding* (NLU) seperti analisis sentimen, parafrase, inferensi bahasa, dan berbagai *task* NLU lainnya. Sehingga untuk dapat melaksanakan domain NLG yang meliputi *text summarization*, *named entity recognition*, dan terutamanya untuk *question answering* diperlukan model bahasa BART (*Bidirectional and Auto-Regressive Transformer*), yang memanfaatkan *auto-denoising encoder*, yang bekerja dengan melakukan *mapping* terhadap dokumen *corrupted* ke representasi aslinya.

Di samping itu, model bahasa BART berhasil mencapai skor 30,6 ROUGE-1, 6,2 ROUGE-2, dan 24,3 ROUGE-L dalam domain *question answering* terhadap dataset ELI5 (Explain Like I am Five) [7] yang ketiga skor ini telah melampaui hasil *state-of-the-art* sebelumnya [7]. Lalu diikuti dengan perkembangan model bahasa mBART yang di-*pretrain* dan di-*fine-tune* dengan korpus *monolingual* dalam berbagai bahasa skala besar, sehingga mampu dalam melakukan domain *machine translation* (MT) terhadap 25 bahasa [8]. Model bahasa mBART sendiri berhasil mencapai skor hingga 12 BLEU terhadap domain *machine translation* ringan dan 5 BLEU untuk tingkat dokumen serta model *unsupervised*. Kemudian model bahasa mBART juga dilakukan *multilingual fine-tuning* sehingga mampu dalam domain *machine translation* terhadap 50 bahasa tanpa berkurangnya performa model baik itu terhadap bahasa dengan sumber daya sedikit, menengah ataupun tinggi. mBART50 yang di-*fine-tuned* secara *multilingual* mampu menghasilkan skor 3.6 BLEU lebih baik dibandingkan dengan *bilingual fine-tuning* [9] sehingga di antara 50 bahasa memungkinkan untuk dilakukan *fine-tuning* dengan dataset TyDiQA [10] untuk tugas *question-answering* dalam bahasa Indonesia yang dapat menjawab pertanyaan secara abstraktif dan kontekstual.

1.2 Rumusan Masalah

Menyesuaikan dengan latar belakang dalam penentuan judul, terdapat rumusan masalah utama dari penelitian di antaranya sebagai berikut.

1. Bagaimana cara penerapan *fine-tuning* dengan *text infilling* untuk tahap *de-*

noising secara optimal terhadap model bahasa BART sehingga menghasilkan model yang dapat diandalkan dalam tugas *question-answering*?

2. Bagaimana performa BLEU dan ROUGE yang dicapai dari model bahasa BART dalam tugas *question-answering* setelah di-*fine-tune* untuk tugas *question-answering*?

1.3 Batasan Permasalahan

Beberapa batasan masalah utama yang ada dari penelitian adalah sebagai berikut.

1. Model bahasa BART yang dirancang untuk tugas *question-answering*.
2. Dataset TyDiQA dalam bahasa Indonesia yang digunakan untuk pelatihan model [10].

1.4 Tujuan Penelitian

Menyesuaikan dengan permasalahan yang telah dijabarkan sebelumnya sehingga didapat beberapa tujuan utama dari penelitian di antaranya sebagai berikut.

1. Menghasilkan model bahasa yang telah di-*fine-tune* sedemikian rupa dari model bahasa BART untuk tugas *question-answering* dengan performa *state-of-the-art* terhadap metrik evaluasi BLEU dan ROUGE yang dapat diandalkan.
2. Mengetahui hasil performa BLEU dan ROUGE dari model bahasa BART yang di-*fine-tuned* dengan *text infilling* pada tahap *denoising* untuk tugas *question-answering*.

1.5 Manfaat Penelitian

Terdapat beberapa manfaat utama dari penelitian di antaranya sebagai berikut.

1. Memberikan wawasan baru bagi para peneliti lapangan di bidang atau industri NLP terhadap kapabilitas beserta kemampuan dari model bahasa BART dalam tugas *question-answering*.

2. Memberikan wawasan baru bagi para peneliti lapangan di bidang atau industri NLP terhadap proses tahapan dari pendekatan *fine-tuning* model bahasa BART untuk tugas *question-answering*.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab satu menjelaskan tentang latar belakang permasalahan rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.

- Bab 2 LANDASAN TEORI
  Bab dua membahas literatur yang melandasi penelitian ini, di antaranya seperti basis model bahasa pendahulu berupa BERT dan BART, kemudian diikuti dengan dataset yang dimanfaatkan yaitu TyDiQA, lalu juga metrik evaluasi mesin berupa BLEU dan ROUGE, serta pustaka penyedia model bahasa yang dikenal dengan HuggingFace.

- Bab 3 METODOLOGI PENELITIAN
  Bab tiga memuat tahapan pendekatan penelitian yang disertai dengan diagram alir.

- Bab 4 HASIL DAN DISKUSI
  Bab empat menjabarkan implementasi sandi, pemrosesan data, pemrosesan model bahasa, *fitting* model bahasa, hingga evaluasi model bahasa.

- Bab 5 KESIMPULAN DAN SARAN
  Bab lima berisi tentang pemaparan kesimpulan hasil penelitian serta saran pengembangan yang dapat dilakukan terhadap penelitian berikutnya.