BAB I

PENDAHULUAN

1.1. Latar Belakang

Kemunculan intelegensia semu memiliki peranan yang cukup besar, dimana intelegensia semu saat ini sudah banyak terlibat dalam kehidupan manusia [1]. Salah satu subbidang pada intelegensia semu yang saat ini terus berkembang pesat adalah *machine learning*. *Machine learning* menggunakan probabilitas dan statistika, tidak seperti metode sistem *rule-based*, yang menggunakan aturan deterministik. Terdapat subbidang di dalam machine learning yang memiliki algoritma yang lebih kompleks, disebut dengan *deep learning*. Salah satu teknik yang dilakukan oleh *deep learning* adalah membangun sebuah kerangka kerja (*framework*) yang memultiplikasi *input* untuk membuat prediksi tentang sifat *input* tersebut. Sistem dapat mengukur seberapa salah prediksi yang ada dengan membandingkannya dengan kebenarannya, dan kemudian menggunakan informasi tersebut untuk memodifikasi algoritmanya, dan hal tersebut yang dilakukan oleh *neural network*. *Neural network* terus mengukur kesalahan dan memodifikasi parameter sehingga mereka dapat mencapai kesalahan sesedikit mungkin.

NLP (*Natural Language Processing*) yang merupakan salah satu perkembangan subbidang dalam intelegensia semu, adalah bagaimana kemampuan komputer dalam memahami bahasa manusia. NLP telah dikembangkan dengan menggunakan pendekatan *deep learning* yang bergantung pada *neural network* dalam

arsitekturnya. Terdapat mekanisme baru yang cukup disruptif dalam pemrosesan sekuensial yaitu *attention*, yang dibawakan dalam model *sequence-to-sequence* (seq2seq) oleh Sutskever pada tahun 2014 [2]. *Attention* sangat mampu menjadi solusi inti dalam isu *context awareness*.

Dalam banyak pendekatan prosedur *training* data dalam *deep learning*, *pretraining* merupakan pendekatan dengan peningkatan terbesar belakangan ini. Dalam *pretraining*, sebuah model terlebih dahulu di-*train* pada *dataset* yang berukuran besar dan *general*. Model yang telah di-*train* yang disebut dengan basis model, kemudian dapat di-*tweak* dengan data dan obyektif yang lebih spesifik. Popularitas *pretraining* meningkat dengan adanya perusahaan seperti Google dan Facebook yang membuat model yang lebih dapat dijangkau dan *open-source*. Terdapat banyak metode *pretrained word embeddings* yang membantu banyak pengembang dapat berbuat lebih, salah satunya yang sedang populer adalah BERT.

BERT merupakan metode *state-of-the-art* dalam pembangunan *language model* dengan pendekatan *deep learning*. BERT menggunakan *transformer*, sebuah mekanisme *attention* yang mempelajari hubungan kontekstual antara kata-kata pada sebuah teks. Secara garis besar, *transformer* mencangkup dua mekanisme terpisah, sebuah *encoder* yang berfungsi untuk membaca teks masukan dan sebuah *decoder* yang memproduksi prediksi untuk pekerjaan tertentu. Dikarenakan tujuan dari BERT adalah untuk menghasilkan sebuah *language model*, hanya mekanisme *encoder*-nya yang dianggap penting. Berlawanan dengan model satu arah atau *directional* yang hanya membaca teks masukan secara sekuensial (*left-to-right* atau *right-to-left*), *encoder* dari

transformer membaca seluruh urutan kata sekaligus. Karakteristik *bidirectional* dari BERT memungkinkan model untuk mempelajari konteks dari sebuah kata berdasarkan lingkungan.

BERT sangat berperan dalam salah satu subtopik dari *Natural Language Processing*, yakni *Natural Language Understanding* (NLU). NLU menginterpretasikan sebuah arti dari bahasa manusia dan mengklasifikasikannya ke intent yang dipahami oleh mesin [3]. Banyak macam pekerjaan dalam NLU yang dapat diselesaikan oleh BERT. Pekerjaan-pekerjaan itu antara lain adalah *Question Answering* (QA), *sentiment analysis*, *machine translation*, *Named Entity Recognition* (NER), dan lainnya.

BERT memberikan peningkatan yang signifikan dalam dunia *machine learning* untuk *Natural Language Processing*. Fakta bahwa BERT hanya membutuhkan pendekatan yang cukup mudah dan memungkinkan *fine-tuning* yang cepat akan memungkinkan berbagai aplikasi yang praktikal di masa depan. Model BERT telah disediakan secara *open-source* oleh Google AI. BERT menyediakan beberapa *pre-trained model*. Salah satu model BERT yang bernama BERT Multilingual merupakan model yang telah dilatih dengan 104 bahasa yang diurutkan dari data Wikipedia terbesar. Namun model pre-trained BERT multilingual masih memiliki banyak limitasi jika digunakan untuk *task* yang hanya menggunakan satu bahasa. BERT Multilingual tidak memiliki deteksi bahasa maupun *selector* bahasa, artinya *tokenizer* perkata dapat bercampur antara bahasa satu dan lainnya. BERT English memiliki ukuran kamus (*vocabulary*) sebesar 28.996 token sedangkan *multilingual model* yang terdiri dari 100 bahasa hanya memiliki 119.547 token untuk seluruh bahasa [4].

Untuk saat ini tidak banyak ditemui basis *language model* berbahasa Indonesia, terlebih lagi belum ada *open-source* basis model bahasa Indonesia dengan teknik BERT sebagai arsitektur model. Dengan membangun model basis yang dilatih hanya dengan *dataset* bahasa Indonesia dengan Teknik BERT, diharapkan model mampu memberikan performa yang cukup baik untuk digunakan atau di-*fine-tuning* untuk pekerjaan yang lebih spesifik seperti klasifikasi, NER, NSP (*Next Sentence Prediction*), dan lainnya.

1.2. Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Apakah *language model* Bahasa BERT memiliki kinerja yang lebih baik dibandingkan dengan BERT pre-trained multilingual?
2. Bagaimana performa *language model* Bahasa BERT dibandingkan dengan BERT *pre-trained* multilingual dalam melakukan *downstream tasks*?

1.3. Tujuan Penelitian

Berdasarkan masalah diatas, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui apakah kinerja *language model* Bahasa BERT lebih baik dibandingkan dengan BERT *pre-trained multilingual*
2. Mengetahui perbandingan performa *language model* Bahasa BERT dengan BERT *pre-trained multilingual* dalam melakukan *downstream tasks*

1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Model Bahasa BERT dapat digunakan sebagai *base model* untuk beberapa pekerjaan yang lebih spesifik
2. Peneliti mendapatkan pengetahuan tentang *Bidirectional Encoder Representations from Transformers* (BERT) yaitu hasil kinerja model Bahasa BERT dalam membangun model bahasa untuk bahasa Indonesia dan algoritma yang digunakan oleh bahasa model.
3. Peneliti mendapatkan pengetahuan tentang cara kerja *transformer*.
4. Dapat menjadi referensi untuk penelitian lebih lanjut dengan topik terkait.

1.5. Batasan Penelitian

Peneliti menetapkan batasan / ruang lingkup penelitian sebagai berikut:

1. Performa model dibandingkan pada performa hasil *fine-tuning* ke dalam *downstream task* berupa klasifikasi teks.
2. Format dataset berlabel yang akan dianalisis adalah teks.
3. Arsitektur model yang digunakan merupakan arsitektur BERT
4. Bahasa yang digunakan adalah bahasa Indonesia
5. Hanya memproses kata-kata yang mengandung kode ASCII atau *American Standard Code for Information Interchange* versi *Printable Characters*.