BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan internet hingga dewasa ini telah banyak membantu manusia dalam berbagai bidang, salah satunya adalah forum *online*. Dalam perkembangannya kini telah ada bermacam-macam situs forum *online* yang dapat dimanfaatkan untuk saling berbagi pengetahuan dan juga berdiskusi. Forum *online* merupakan platform yang digunakan oleh pengguna dari berbagai belahan dunia untuk mengumpulkan, berbagi informasi, dan berdiskusi antar pengguna mengenai minat dan kesamaan topik [1]. Pengguna dalam forum *online* dapat menanyakan pertanyaan mengenai sebuah topik, kemudian pengguna lain yang ahli dalam topik tersebut dapat menjawab pertanyaan. Sebuah pertanyaan yang diikuti kumpulan jawaban dari berbagai pengguna disebut *thread*. Namun dikarenakan setiap pengguna memiliki cara berbeda dalam mengajukan pertanyaan, terkadang pengguna menanyakan pertanyaan yang secara semantik sama dengan pertanyaan yang telah diajukan pada *thread* lain [2].

Quora adalah salah satu contoh dari forum *online* yang ada dan juga populer digunakan. Sebagai forum tanya jawab, Quora menghadapi masalah berupa pertanyaan duplikat pada seluruh *thread*nya. Untuk mengatasi masalah pertanyaan duplikat ini, pertanyaan yang mirip akan dipindahkan ke *thread* yang sama. Sejauh ini, pemindahan pertanyaan dilakukan secara manual oleh moderator dan melalui laporan pengguna [3].

Untuk itulah peran *Natural Language Processing* atau biasa disingkat NLP diperlukan untuk melakukan pendeteksian pertanyaan yang memiliki kemiripan secara otomatis. Salah satu terapan NLP sendiri adalah *semantic similarity*, yang mana dapat diterapkan secara luas pada berbagai kasus termasuk kasus pertanyaan duplikat ini. *Semantic similarity* merupakan metode untuk mengukur kesamaan/keterkaitan antara pasangan kata secara semantik. Didasari oleh faktor di atas, maka dalam penelitian ini penulis akan mengimplementasikan metode yang baru-baru ini dikembangkan dalam bidang NLP yaitu *Bidirectional Encoder Representations from Transformers* (BERT) untuk diuji performa serta keakuratannya dalam mengukur kesamaan pertanyaan pengguna.

BERT digunakan karena meskipun kini BERT telah mencapai berbagai hasil yang menakjubkan pada beberapa tugas NLP dan mengungguli metode lain seperti Word2Vec, Glove, ELMo, dan sebagainya, namun nilai potensialnya masih belum dieksplorasi secara penuh khususnya oleh peneliti di Indonesia. Diharapkan setelah penelitian ini dilakukan dan telah diketahui performa dari BERT nantinya akan dapat diimplementasikan oleh praktisi di Indonesia untuk dapat menciptakan sistem yang lebih cerdas, dan juga tidak menutup kemungkinan agar penelitian ini dapat dikembangkan lebih lanjut oleh peneliti lain sehingga ilmu pengetahuan di Indonesia pada bidang NLP dan *Deep Learning* terus berjalan dan berkembang.

1.2. Rumusan Masalah

Dari latar belakang yang dipaparkan di atas didapatkan rumusan masalah yaitu berapa nilai akurasi yang diberikan oleh metode BERT dalam melakukan prediksi kesamaan pertanyaan?

1.3. Batasan Masalah

Adapun batasan masalah yang akan dirancang dalam penelitian ini adalah:

1. Kode program serta *library* yang diimplementasikan bertujuan hanya untuk mengetahui kemampuan serta tingkat akurasi metode BERT dalam memberikan nilai *semantic similarity*.
2. Data pertanyaan akan diambil dari situs Quora.com dengan topik/tema seputar "Teknologi Informasi dan Komunikasi".
3. Bahasa pertanyaan yang digunakan adalah Bahasa Indonesia.
4. Hasil yang diberikan berupa label prediksi kemiripan dan nilai probabilitas prediksi (dalam persen).

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dilakukannya penelitian ini adalah untuk mengimplementasikan metode BERT sehingga mampu diketahui performa serta akurasi metode tersebut dalam memberikan nilai kemiripan makna kalimat antara teks pertanyaan asli dengan pertanyaan pembanding.

1.5. Manfaat Penelitian

Adapun manfaat penelitian antara lain :

1. Untuk mengetahui implementasi BERT dalam melakukan perhitungan *sentence similarity*.
2. Untuk mengetahui performa dan tingkat akurasi yang dihasilkan dari metode BERT yang diimplementasikan.

1.6. Metodologi Penelitian

Metodologi yang digunakan dalam proses pengerjaan penelitian ini adalah sebagai berikut :

1.6.1 Pengumpulan *Dataset*

Data yang digunakan di dalam penelitian ini diambil dari situs forum tanya jawab Quora dengan teknik *crawling*. Data yang akan digunakan sebagai pertanyaan asli adalah sebanyak 12.000 data. Seluruh data yang didapatkan merupakan data berupa teks yang disimpan dalam format .csv. Nantinya *dataset* akan dipisah menjadi tiga bagian yaitu data *train* yang akan digunakan dalam proses *training*, data *dev* atau *validation set* yang tidak dipelajari oleh model namun digunakan untuk memilih model yang terbaik, dan data *test* yang akan digunakan dalam proses pengujian.

1.6.2 Pengisian Data Pertanyaan Uji dan *Labelling*

Sebanyak 12.000 pertanyaan yang telah didapatkan sebelumnya adalah pertanyaan asli dari situs Quora. Sedangkan untuk pertanyaan uji ditulis secara manual oleh peneliti sehingga menghasilkan pasangan pertanyaan. Selain itu pasangan pertanyaan juga diberi *question id* dan juga label seperti 0 yang menandakan tidak duplikat, dan 1 yang menandakan pertanyaan adalah duplikat. Nantinya label ini akan digunakan sebagai validasi antara penilaian sistem dengan penilaian sebenarnya.

1.6.3 *Preprocessing* Data

Data yang telah disiapkan selanjutnya dilakukan proses *preprocessing* hanya melalui tahap *tokenizing* dan *case folding*. Ini dimaksudkan untuk membersihkan dan mempersiapkan data sebelum dapat dilatih dan diuji.

1.6.4 Tahap *Fine-Tuning*

Ini adalah tahap dimana dilakukan penyesuaian *model* dengan tugas *semantic similarty* yang akan dilakukan. *Model* (atau biasa disebut data *pre-train)* sendiri berisikan konfigurasi, *vocabulary*, dan vektor/*embedding* kata. Dalam penelitian ini *model* bahasa Indonesia sudah dibuat oleh tim IndoNLU dengan menggunakan *corpus* Indo4B yang berisikan artikel dari berbagai media massa *online*, sosial media, dan juga artikel Wikipedia bahasa Indonesia. Kemudian dalam penerapannya hanya perlu menggunakan data

pre-train yang telah dibuat untuk melakukan proses *fine-tuning* (penyesuaian) *model* terhadap data *train*.

1.6.5 Pengujian

Ini merupakan tahap pengujian dimana dilakukan prediksi label yang dilakukan oleh algoritma terhadap data *test*. Pasangan pertanyaan akan dibaca probabilitas kesamaan kalimatnya secara semantik menggunakan algoritma *logistic regression*.

1.6.6 Evaluasi

Pada tahap ini penulis melakukan perhitungan akurasi (dalam bentuk persen) dari prediksi yang telah dilakukan, dan selanjutnya melihat ketepatan BERT dalam melakukan prediksi menggunakan algoritma *confusion matrix*. Selain itu dilakukan juga analisa mengenai kelebihan dan kekurangan metode yang telah diimplementasikan, serta pengambilan kesimpulan terhadap hasil yang telah didapat.

1.7. Sistematika Penulisan

Untuk mengatahui uraian singkat yang memuat gambaran singkat secara keseluruhan isi masing-masing bab, maka dibuat sitematika sebagai berikut :

BAB I PENDAHULUAN

Berisi tentang latar belakang penelitian, rumusan masalah, manfaat penelitian, batasan masalah, tujuan penelitian,

metodologi, sistematika penulisan penelitian, serta rencana jadwal kegiatan penelitian.

BAB II LANDASAN TEORI

Bab ini berisi tentang teori-teori yang digunakan serta konsep dasar yang menjadi landasan dari penelitian ini yang berasal dari berbagai sumber seperti jurnal dan *proceeding*.

BAB III METODE PENELITIAN

Bab ini berisi gambaran umum penelitian, alat dan bahan penelitian, serta langkah-langkah penelitian.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Bab ini akan membahas mengenai proses implementasi dan pengembangan yang dilakukan secara menyeluruh serta pembahasan pada penelitian yang dikaji.

BAB V PENUTUP

Bab ini berisi kesimpulan dari penelitian yang telah dilakukan serta saran kedepannya untuk memperbaiki kekurangan yang ada penelitian.