Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Perpustakaan Univeristas Multimedia Nusantara

BAB V

SIMPULAN DAN SARAN

5.1. Simpulan

Sesuai dengan rumusan masalah yang ditetapkan di awal penelitian berlangsung, dapat disimpulkan bahwa Bahasa BERT mampu meraih performa yang lebih baik dibandingkan dengan BERT *pre-trained* multilingual. Setelah dilakukan validasi silang pada *dataset* yang mayoritas terdiri dari Bahasa Indonesia, performa AUC setelah *softmax* di Bahasa BERT dapat melampaui BERT Multilingual sebesar 0.58%. Pada klasifikasi teks *real-world case* juga didapati performa AUC setelah *softmax* di Bahasa BERT dapat melampaui BERT Multilingual sebesar 3.56%. Ukuran kosakata dari kedua model memiliki selisih yang jauh. BERT Multilingual memiliki 119.547 kosakata dan Bahasa BERT yang hanya memiliki 31.743 kosakata, yang juga membuat Bahasa BERT pada proses *fine-tuning* membutuhkan durasi yang lebih singkat dibandingkan dengan BERT Multilingual dengan rata-rata kecepatan 1.94 kali lebih cepat.

Performa dari Bahasa BERT dalam melakukan *downstream task* yang di mana *task* tersebut merupakan tugas yang menitikberatkan bahasa Indonesia dibandingkan dengan bahasa lain, dapat melampaui BERT Multilingual yang juga dilatih dengan beberapa kosakata dalam Bahasa Indonesia. Waktu yang lebih singkat untuk proses *fine-tuning* juga membuat Bahasa BERT lebih efektif dan efisien dalam penggunaannya dibandingkan dengan BERT Multilingual.

5.2. Saran

Dari rangkaian proses yang telah dilalui, terdapat beberapa hal yang dapat dijadikan saran terkait penelitian ini antara lain :

1. Penelitian ini dapat dilanjutkan dengan menambahkan *dataset* tanpa label dalam proses *training* untuk *pretrained* Bahasa BERT agar sumber korpus dapat lebih banyak lagi dan memperluas pengetahuan model;
2. Penelitian ini juga dapat dicoba menggunakan metode *tokenizer* lain seperti *tokenizer* WordPiece yang meng-*tokenize* tiab subkata, sedangkan pada penelitian ini menggunakan SentencePiece yang bersifat *open source*, dengan hal tersebut bisa dibandingkan metode yang lebih baik untuk *tokenize*;
3. Penelitian ini dapat dilakukan dengan *resource* yang memadai agar penelitian dapat dilakukan dalam beberapa kali percobaan untuk mendapatkan hasil yang optimal;
4. Dikarenakannya pada penelitian dapat dilihat bahwa *downstream task* standar itu penting dalam dunia penelitian, dapat dilakukan peneltian pada penetapan *downstream task* standar yang juga menggunakan model pada penelitian ini sebagai salah satu tolak ukur.