ABSTRAK

IndoBERT: Transformer-based Model for Indonesian Language Understanding

By

Sarah Lintang Sariwening
18/433796/PPA/05611

Model bahasa berbasis deep learning yang telah dilatih pada *unannotated* teks berjumlah besar telah dikembangkan untuk efisiensi *transfer learning* pada pemrosesan bahasa alami. Model berbasis Transformer seperti BERT sangat populer akhir-akhir ini karena performa yang sangat baik. Namun, kebanyakan model tersebut hanya berfokus pada bahasa inggris, dan hanya sedikit yang berfokus pada bahasa dengan resource yang terbatas seperti bahasa Indonesia. Oleh karena itu, penelitian ini bertujuan membangun model monolingual BERT untuk bahasa Indonesia (IndoBERT), yang memperlihatkan keunggulan performanya dibanding dengan model lainnya atau model Multilingual BERT (M-BERT).

Pada penelitian ini kami membangun IndoBERT dari awal. Model ini secara konstan mengungguli multilingual BERT model pada aplikasi pemrosesan bahasa alami berbasis bahasa Indonesia seperti sentimen analisis dan peringkasan teks.

Kata kunsi: Data, NLP, Language Modeling, Text Summarization, BERT

ABSTRACT

IndoBERT: Transformer-based Model for Indonesian Language Understanding

By

Sarah Lintang Sariwening
18/433796/PPA/05611

Deep learning-based language models pre-trained on large unannotated text corpora have been developed to allow efficient transfer learning for natural language processing. A recent approach, Transformer-based models such as BERT, has become increasingly popular due to their state-of-the-art performance. However, most work of these models are usually focused on English, leaving other languages to multilingual models with limited resources. This paper proposes a monolingual BERT for the Indonesian language (IndoBERT), which shows its state-of-the-art performance compared to other architectures and Multilingual BERT (M-BERT) models.

We built IndoBERT from scratch. This model consistently outperforms the multilingual BERT model on downstream NLP tasks such as the Sentiment Analysis and Summarization task.

Keywords: Data, NLP, Language Modeling, Text Summarization, BERT