Pengenalan Entitas Bernama pada Twit Berbahasa Indonesia Menggunakan Model Pre-Trained BERT = BERT Pre-Trained Language Model for Named Entity Recognition on Indonesian Tweets

Muhammad Anwar Farihin, author

Abstrak

Pengenalan Entitas Bernama (NER) telah diteliti cukup dalam, khususnya pada korpus berbahasa Inggris. Namun, penelitian NER pada korpus twit berbahasa Indonesia masih sangat sedikit karena minimnya dataset yang tersedia secara publik. BERT sebagai salah satu model state-of-the-art pada permasalahan NER belum diimplementasikan pada korpus twit berbahasa Indonesia. Kontribusi kami pada penelitian ini adalah mengembangkan dataset NER baru pada korpus twit berbahasa Indonesia sebanyak 7.426 twit, serta melakukan eksperimen pada model CRF dan BERT pada dataset tersebut. Pada akhirnya, model terbaik pada penelitian ini menghasilkan nilai F1 72,35% pada evaluasi tingkat token, serta nilai F1 79,27% (partial match) dan 75,40% (exact match) pada evaluasi tingkat entitas.

.....Named Entity Recognition (NER) has been extensively researched, primarily for understanding the English corpus. However, there has been very little NER research for understanding Indonesian-language tweet corpus due to the lack of publicly available datasets. As one of the state-of-the-art models in NER, BERT has not yet been implemented in the Indonesian-language tweet corpus. Our contribution to this research is to develop a new NER dataset on the corpus of 7.426 Indonesian-language tweets and to conduct experiments on the CRF and BERT models on the dataset. In the end, the best model of this research resulted in an F1 score of 72,35% at the token level evaluation and an F1 score of 79,27% (partial match) and 75,40% (exact match) at the entity level evaluation.