Normalisasi Kata pada Teks Twitter Berbahasa Campuran Indonesia-Inggris menggunakan UFAL ByT5 = Text Normalization on Indonesian-English Code-Mixed Twitter Text using UFAL ByT5

Rafi Dwi Rizqullah, author

Abstrak

Media sosial telah berkembang pesat dalam masyarakat dunia. Tak terkecuali Twitter yang mendapatkan peningkatan baik dalam jumlah pengguna maupun konten yang dibuat. Namun, Twitter memiliki batasan karakter dalam satu tweet yang menyebabkan perubahan pada pola penulisan para penggunanya. Pengguna Twitter mulai memodifikasi penulisan dengan kata baku menjadi kata tidak baku, salah satunya dengan menggunakan bahasa campuran. Untuk keperluan analisis tweet, normalisasi teks diperlukan untuk mengubah kata tidak baku menjadi baku untuk memudahkan analisis. State-of-the-art pada normalisasi teks Twitter berbahasa campuran Indonesia dan Inggris saat ini adalah model statistical machine translation (SMT), namun model SMT masih memiliki kelemahan pada beberapa jenis perubahan kata. Penelitian ini berfokus pada normalisasi teks Twitter Indonesia berbahasa campuran Indonesia dan Inggris dengan menggunakan salah satu model transformer yaitu UFAL ByT5. Terdapat dua model UFAL ByT5 yang digunakan masing-masing untuk bahasa Indonesia serta bahasa Inggris. Hasil penelitian menunjukkan model UFAL ByT5 unggul dalam normalisasi teks dibandingkan model SMT, dengan selisih nilai BLEU 0,88 persen lebih besar.

.....Social media has been grown rapidly in the global community. It also includes Twitter, which is getting increase in both users and content created. However, Twitter has character limit in one tweet which causes changes to the writing patterns of its users. Twitter users began to modify their writing from using formal words into non-formal words, one of which was using code-mixed language. For tweet analysis purposes, text normalization is required to transform non-formal words into formal ones to help analysis process. The recent state-of-the-art for Indonesian-English code-mixed Twitter text normalization is with statistical machine translation (SMT) models, however the SMT model still has weakness in word recognition. This research focuses on the Indonesian and English code-mixed Twitter text normalization using one of transformer model which is UFAL ByT5. There are two UFAL ByT5 models that were used, each of them are for Indonesian and English language. Research result shows that UFAL ByT5 model outperform SMT model on text normalization by 0.88 percent of BLEU score in difference.