INTISARI

Pemanfaatan Normalisasi Teks pada Augmentasi Data Teks User-Generated Content Berbahasa Indonesia

Oleh

Faturahman Yudanto
22/498827/PPA/06325

User-generated content sering digunakan pada ranah pemrosesan bahasa natural, terutama pada kasus klasifikasi teks. Tantangan memproses teks *user-generated content* berbahasa Indonesia adalah terkait dengan gaya penulisan yang cenderung informal. Selain itu, terbatasnya dataset klasifikasi teks dalam Bahasa Indonesia untuk kasus tertentu membuat pengembangan model klasifikasi teks membutuhkan waktu yang lama. Teknik augmentasi data pada teks *user-generated content* berbahasa Indonesia, yang mana dapat mengatasi keterbatasan data, masih belum mengatasi keragaman yang ada seperti teks berbentuk slang dan singkatan. Di sisi lain, proses normalisasi teks bentuk slang dan singkatan juga akan menambah beban komputasi saat dilakukan pada proses inferensi.

Penelitian ini melibatkan pengumpulan dataset, pembuatan model normalisasi teks, augmentasi data teks, pelatihan model klasifikasi teks, dan evaluasi performa model. Model normalisasi teks dikembangkan menggunakan *deep learning* berbasis *sequence to sequence* yang selanjutnya digunakan untuk augmentasi data pada data latih kasus klasifikasi teks. Augmentasi data dilakukan pada pelatihan model klasifikasi teks menggunakan model Fasttext-Bi-LSTM dan *pretrained model* IndoBERTweet, masing-masing dengan skenario augmentasi data teks secara *offline* dan *online* dengan modifikasi.

Hasil penelitian ini didapatkan bahwa normalisasi teks untuk augmentasi data teks dapat meningkatkan performa model klasifikasi teks. Pada dataset klasifikasi emosi, augmentasi data dengan normalisasi teks secara online dengan probabilitas 0,7 meningkatkan nilai *F1 score* tertinggi dari sebelumnya 0,7701 menjadi 0,7827. Sementara itu pada dataset PRDECT-ID, augmentasi data dengan normalisasi teks secara offline meningkatkan nilai *F1 score* tertinggi dari 0,6406 menjadi 0,7002 dengan penambahan data sebesar 100%. Kedua nilai *F1 score* terbaik didapatkan oleh model IndoBERTweet.

Kata-kata kunci : *User-generated content*, Augmentasi Data Teks, Normalisasi Teks

ABSTRACT

The Use Of Text Normalization On Indonesian User-Generated Content Text Data Augmentation

By

Faturahman Yudanto
22/498827/PPA/06325

User-generated content is widely used in natural language processing tasks, especially in text classification. One of the challenges in processing Indonesian user-generated content is its informality. Moreover, the data availability for certain use cases also still remains a challenge, and it makes the development process of text classification models take a long time. Although text data augmentation can be used to tackle the data availability problem, current text data augmentation techniques are still unable to consider the text informality, which is sometimes written in a slang word or abbreviation. On the other hand, the normalization of slang and abbreviated words also increases the computational time when it is done in the inference process. Therefore, there is a possibility of the development of new text data augmentation using text normalization.

This research involves collecting dataset, developing text normalization model, using the model for text data augmentation, and evaluating the effect of the data augmentation method on text classification tasks. Text normalization model is developed using character sequence-to-sequence approach and then used for text data augmentation in text classification tasks. A Fasttext-BiLSTM model and a pretrained IndoBERTweet model are used for the text classification tasks. Both models are trained on both offline and modified online text data augmentation scenarios.

The result indicates that the use of text normalization for data augmentation is able to improve the text classification model performance in certain scenarios. In the emotion classification dataset, online data augmentation with a replacement probability of 0.7 yielded the most favorable outcome, demonstrating an improvement of F1 Score from 0.7701 to 0.7827. Conversely, in the PRDECT-ID dataset, the optimal results were achieved through offline data augmentation with a 100% proportion of data increment, resulting in an increase of F1 Score from 0.6406 to 0.7002. Both best F1 Score are produced by IndoBERTweet model.

Keywords : User-generated content, Text Data Augmentation, Text Normalization