Normalisasi teks code-mixed bahasa Indonesia-Inggris pada data twitter dan analisis pengaruhnya untuk klasifikasi emosi = Code-mixed text normalization on Indonesian-English language on twitter data and the analysis of its effect on emotion classification.

Ajmal Kurnia, author

Abstrak

Code-mixing adalah sebuah fenomena pengunaan dua atau lebih bahasa dalam suatu percakapan. Fenomena ini semakin banyak digunakan oleh pengguna internet Indonesia yang mencampur bahasa Indonesia-Inggris. Normalisasi teks code-mixed ke dalam satu bahasa perlu dilakukan agar kata-kata yang ditulis dalam bahasa lain dalam teks tersebut dapat diproses dengan efektif dan efisien. Penelitian ini melakukan normalisasi teks code-mixed pada bahasa Indonesia-Inggris dengan menerjemahkan teks ke dalam bahasa Indonesia. Penulis melakukan pengembangan pada pipeline normalisasi code-mixed dari penelitian sebelumnya sebagai berikut: melakukan rekayasa fitur pada proses identifikasi bahasa, menggunakan kombinasi ruleset dan penerjemahan mesin pada proses normalisasi slang, dan menambahkan konteks pada proses Matrix Language Frame (MLF) pada proses penerjemahan. Hasil eksperimen menunjukkan bahwa model identifikasi bahasa yang dibuat dapat meningkatkan nilai F1-score 4,26%. Model normalisasi slang yang dibuat meningkatkan nilai BLEU hingga 25,22% lebih tinggi dan menunrunkan nilai WER 62,49%. Terakhir, proses penerjemahan yang dilakukan pada penelitian ini berhasil memperoleh nilai BLEU 2,5% lebih tinggi dan metrik WER 8,84% lebih rendah dibandingkan dengan baseline. Hasil ini sejalan dengan hasil eksperimen keseluruhan pipeline. Berdasarkan hasil eksperimen keseluruhan pipeline yang dibuat oleh penulis dapat meningkatkan secara signifikan performa BLEU hingga 32,11% dan menurunkan nilai WER hingga 33,82% lebih rendah dibandingkan dengan metode baseline. Selanjutnya, penelitian ini juga menganalisis pengaruh dari proses normalisasi teks code-mixed untuk klasifikasi emosi. Proses normalisasi teks code-mixed terbukti dapat meningkatkan performa sistem klasifikasi emosi hingga 12,45% untuk nilai F1-score dibandingkan dengan hanya melakukan tokenisasi dan meningkatkan nilai F1-score hingga 6,24% dibandingkan dengan metode preproses sederhana yang umum digunakan. Hal ini menunjukkan bahwa normalisasi teks code-mixed memiliki pengaruh positif terhadap efektifitas pemrosesan teks, sehingga normalisasi ini penting untuk dilakukan pada task yang menggunakan data code-mixed.

.....

Code-mixing is the mixing of two or more languages in a conversation. The usage of code-mixing has increased in recent years among Indonesian internet users that often mixed Indonesian language with English. Normalization of code-mixed text has to be applied to translate code-mixed text so that the text can be processed effectively and efficiently. This research performed code-mixed text normalization on Indonesian-English text by translating the text to Indonesian language. Author improves existing normalization pipeline from previous research by: (1) feature engineering on language identification, (2) using combination of ruleset and machine translation approach on slang normalization, and (3) adding some context on matrix language frame that used on translation process. Experiment result shows language identification model that developed in this research is able to improve F1-score by 4,26%. Slang normalization model from this research is able to improve BLEU score by 25,22% and lower WER score by

62,49%. Lastly, translation process on this research is able to improve BLEU score by 2,5% and lower WER score by 8,84% compared to baseline. Experiment results on the entire normalization pipeline shows similar results. The result shows the new pipeline is able to significantly improves previous pipeline by 32,11% on BLEU metric and reduces WER by 33,82% compared to baseline normalization system. This research also tried to analyze the effect of code-mixed text normalization process on emotion classification. Code-mixed text normalization is able to improve evaluation result of emotion classification model by 12,45% on F1-score compared to tokenization only preprocessing data and 6,24% compared to common text preprocessing method. This result shows that the code-mixed text normalization has positive effect to text processing and also shows the importance to perform this normalization when using code-mixed data.