

ABSTRACT

Twitter is one of the platforms on the internet where people go to pour their heart out. There are a lot of personal, everyday-things that are posted. Explicitly these Tweets contain emotions of the writer that can show what state of mind they're in. These tweets are essential for emotion research in pursue of gaining a better understanding of people. By understanding people's emotional states, people can unlock the keys to better managing their emotional responses and in the long-term improving their mental health and overall well-being. But since people are capable of experiencing numerous emotions at once, talking about human emotion is a very complicated subject. This research intends to create a data augmentation approach that can be applied to multi-label Indonesian tweets, as there is currently little to no such dataset. The proposed augmentation method utilizes Bidirectional Encoder Representations Transformers (BERT) to swap words on a tweet to their synonym. Two classifiers of Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) then try to classify the augmented dataset. The effectiveness of augmentation is then evaluated by comparing the performance of classifiers trained on the original dataset with that of classifiers trained on the augmented dataset. Another comparison is also done to see the augmentation effect on different embedding techniques used. Evaluation result shows proposed augmentation improve classifier performance up to 3.36% for SVM micro F1 and up to 3.84% for LSTM micro F1. Proposed augmentation can also work with FastText embedding pushing the performance even higher to 4.52% improvement in micro F1 for LSTM.

INTISARI

Twitter adalah salah satu platform di internet tempat orang mencurahkan isi hatinya. Ada banyak hal pribadi sehari-hari yang diposting. Secara eksplisit Tweet ini mengandung emosi penulis yang dapat menunjukkan keadaan pikiran mereka. Tweet ini sangat penting untuk penelitian emosi dalam upaya mendapatkan pemahaman yang lebih baik tentang orang. Dengan memahami keadaan emosional orang, orang dapat membuka kunci untuk mengelola respons emosional mereka dengan lebih baik dan dalam jangka panjang meningkatkan kesehatan mental dan kesejahteraan mereka secara keseluruhan. Tetapi karena orang mampu mengalami banyak emosi sekaligus, berbicara tentang emosi manusia adalah topik yang sangat rumit. Penelitian ini bertujuan untuk membuat pendekatan augmentasi data yang dapat diterapkan pada kicauan multi-label Indonesia, karena saat ini dataset seperti itu masih sedikit atau tidak ada sama sekali. Metode augmentasi yang diusulkan menggunakan Bidirectional Encoder Representations Transformers (BERT) untuk menukar kata pada tweet ke sinonimnya. Dua pengklasifikasi Support Vector Machine (SVM) dan Long Short-Term Memory (LSTM) kemudian mencoba mengklasifikasikan kumpulan data yang ditambah. Efektivitas augmentasi kemudian dievaluasi dengan membandingkan kinerja pengklasifikasi yang dilatih pada kumpulan data asli dengan pengklasifikasi yang dilatih pada kumpulan data yang ditambah. Perbandingan lain juga dilakukan untuk melihat efek augmentasi pada berbagai teknik embedding yang digunakan. Hasil evaluasi menunjukkan augmentasi yang diusulkan meningkatkan kinerja classifier hingga 3,36% untuk SVM mikro F1 dan hingga 3,84% untuk LSTM mikro F1. Augmentasi yang diusulkan juga dapat bekerja dengan penyematan FastText yang mendorong kinerja lebih tinggi hingga peningkatan 4,52% dalam mikro F1 untuk LSTM.