Identifikasi Ujaran Kebencian dan Ujaran Kasar pada Twit Berbahasa Campuran Indonesia-Jawa dengan Pre-Trained Language Model Berbasis BERT = Hate-Speech and Abusive Language Identification on Code-Mixed Indonesian and Javanese Language Tweets Using BERT-based Pre-trained Language Model

Alif Mahardhika, author

Abstrak

Ujaran kasar dan ujaran kebencian telah menjadi fenomena yang banyak ditemukan di media sosial. Penyalahgunaan kebebasan berpendapat ini berpotensi memicu terjadinya konflik dan ketidakstabilan sosial dikalangan masyarakat, baik dalam interaksi sosial secara digital maupun secara fisik. Diperlukan upaya identifikasi ujaran kasar dan ujaran kebencian secara otomatis, akurat, dan efisien untuk mempermudah penegakkan hukum oleh pihak berwenang. Penelitian pada skripsi ini melakukan perbandingan performa klasifikasi ujaran kasar dan ujaran kebencian pada data teks mixed-coded berbahasa Indonesia-Jawa, menggunakan model klasifikasi berbasis BERT. Eksperimen perbandingan dilakukan dengan membandingkan pre-trained model berbasis BERT dengan berbagai arsitektur dan jenis berbeda, yaitu BERT (dengan arsitektur base dan large), RoBERTa (arsitektur base), dan DistilBERT (arsitektur base). Untuk mengatasi keterbatasan mesin dalam memahami teks mixed-coded, penelitian ini dirancang dalam dua skenario yang membandingkan performa klasifikasi pada teks mixed-coded Indonesia-Jawa dan teks mixed coded yang diterjemahkan ke Bahasa Indonesia. Hasil terbaik berdasarkan F1-Score didapatkan pada klasifikasi menggunakan model berbasis BERT dengan nama IndoBERT-large-p2 pada kedua skenario, dengan F1-Score 78,86% pada skenario tanpa proses translasi, dan F1-Score 77,22% pada skenario dengan proses translasi ke Bahasa Indonesia.

.....Hateful and abusive speech has become a phenomenon that becomes common in social media. This abuse of freedom of speech presents significant risk of starting social conflicts, be it in the form of digital or physical social interactions. An accurate, efficient, and automated hate speech and abusive language identification effort needs to be developed to help authorities address this problem properly. This research conducts a comparison on hate speech and abusive language identification using several BERT-based language models. The comparisons are made using a variety of BERT-based language models with different types and architecture, including BERT (base and large architecture), RoBERTa (base architecture), and DistilBERT (base architecture). To address the mixed-coded nature of social media texts, this research was conducted under two different scenario that compares the classification performance using a mixed-coded Indonesian-Javanese text and texts that have been translated to Indonesian. The best classification output was measured using F1-Score, with a BERT-based model named IndoBERT-large-p2 outscoring the other BERT-based models in both scenario, scoring an F1-Score of 78.86% in untranslated scenario, and 72.22% F1-Score on the Indonesian-translated scenario.