Universitas Indonesia Library >> UI - Skripsi Membership

Pembangunan Data dan Model Analisis Emosi Fine-Grained pada Teks Media Sosial Berbahasa Indonesia = Fine-Grained Emotion Analysis on Indonesian Social Media Text: Dataset and Models

Gilang Catur Yudishtira, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920533026&lokasi=lokal

---

Abstrak

Saat ini, dataset yang tersedia untuk melakukan analisis emosi di Indonesia masih terbatas, baik dari segi jumlah data, cakupan emosi, serta sumbernya. Pada penelitian ini, peneliti membangun dataset besar untuk tugas analisis emosi pada data teks berbahasa Indonesia, di mana dataset ini dikumpulkan dari berbagai domain dan sumber. Dataset ini mengandung 33 ribu teks, yang terdiri dari tweet yang dikumpulkan dari Twitter, serta komentar unggahan yang dikumpulkan dari Instagram dan Youtube. Domain yang dicakup pada dataset ini adalah domain olahraga, hiburan, dan life chapter. Dataset ini dianotasi oleh 36 annotator dengan label emosi fine-grained secara multi-label, di mana label emosi yang digunakan ini merupakan hasil dari taksonomi emosi baru yang diusulkan oleh peneliti. Pada penelitian ini, peneliti mengusulkan taksonomi emosi baru yang terdiri dari 44 fine-grained emotion, yang dikelompokkan ke dalam 6 basic emotion. Selain itu, peneliti juga membangun baseline model untuk melakukan analisis emosi. Didapatkan dua baseline model, yaitu hasil fine-tuning IndoBERT dengan f1-score micro tertinggi sebesar 0.3786, dan model hierarchical logistic regression dengan exact match ratio tertinggi sebesar 0.2904. Kedua baseline model tersebut juga dievaluasi di lintas domain untuk dilihat seberapa general dan robust model yang telah dibangun.

.....Currently, no research in Indonesia utilises fine-grained emotion for emotion analysis. In addition, the available datasets for analysing emotions still need to be improved in terms of the amount of data, the range of emotions, and their sources. In this study, researchers built a large dataset for analysing emotion. This dataset contains 33k texts, consisting of tweets collected from Twitter and comments collected from Instagram and Youtube posts. The domains covered in this dataset are sports, entertainment, and life chapter. Thirty-six annotators annotated this dataset with fine-grained emotion labels and a multi-label scheme, where the emotion labels resulted from a new emotion taxonomy proposed by the researcher. In this study, the researchers propose a new emotion taxonomy consisting of 44 fine-grained emotions which are grouped into six basic emotions. Two baseline models were obtained, the first one is the fine-tuned IndoBERT model, which achieved the highest f1-score micro of 0.3786, and the second one is hierarchical logistic regression model, which achieved the highest exact match ratio of 0.2904. Both baseline models were also evaluated to determine their cross-domain applicability. The dataset and baseline models that are produced in this study are expected to be valuable resources for future research purposes.