TRANSFORMER-BASED STRESS DETECTION IN DEVELOPER DISCUSSIONS USING STACK OVERFLOW DATA

Pinnalawaththage Dona Githmi Hashara[1*], Shanmuganathan Vasanthapriyan[2], Thushani Dhananjalee Rubasinghe[3]

[1]Department of Computing and Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.

[2]Department of Data Science, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.

[3]Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka.

*Corresponding author: pdghashara@std.appsc.sab.ac.lk

Abstract

Developer stress can affect productivity, software quality, and developer well-being. Although stress detection has been explored on general social media platforms, identifying stress-related linguistic signals in technical discussions remains relatively underexplored. This study investigates transformer-based classification of stress-related content in Stack Overflow posts. A dataset of 6,992 manually annotated posts was developed through an annotation process involving three independent annotators. Inter-annotator agreement was assessed using a pilot sample of 300 posts, yielding a Krippendorff's alpha of 0.86. The annotation scheme categorizes posts into three groups: stress-related posts, neutral posts containing stress-related keywords, and neutral posts without stress-related keywords. For model training, the two neutral categories were merged into a single class, resulting in a binary classification task. To examine the influence of training data scale, experiments used training subsets of 1,500, 2,500, and 3,500 samples. BERT-base, RoBERTa-base, and DistilBERT-base were fine-tuned and evaluated using accuracy, precision, recall, and F1-score. RoBERTa-base achieved 87.65% accuracy, an F1-score of 82.87% for the stress class, and a stress-class recall of 89.50% using 3,500 training samples. A TF-IDF-based Logistic Regression baseline achieved 74.83% accuracy and 65.26% macro-F1. The findings highlight the potential of contextual language representations for identifying stress-related linguistic signals in developer discussions and underscore the importance of annotation reliability, training data scale, and linguistic ambiguity in text classification for developer well-being research.

Keywords: Developer stress, Stack Overflow, Transformer models, Text classification, Developer well-being