Evaluating Pre-Trained Models for User Feedback Analysis in Software Engineering: A Study on Classification of App-Reviews

Mohammad Abdul Hadi
Department of Computer Science
University of British Columbia
Kelowna, Canada
mohammad.hadi@ubc.ca

Fatemeh H. Fard
Department of Computer Science
University of British Columbia
Kelowna, Canada
fatemeh.fard@ubc.ca; ORCID: 0000-0002-4505-6257

I. Extended Abstract

Context: Mobile application users' feedback has been studied extensively in Software Engineering, including requirement engineering, release planning, software maintenance, change-file localization, and testing. Most of the research focused on developing classifiers for various purposes. However, supervised classification requires a lot of manually labeled data, and with introducing new classes or new platforms (e.g., App Store, Twitter, Google Play), new labeled data and models are required. Recently, Pre-trained Neural Language Models (PLMs) have gained much attention in Natural Language Processing (NLP) and Software Engineering (SE). Though they have been used extensively in NLP for text classification and many other tasks, such as question answering, their applicability has yet to be explored for app review classification. Specifically, because PLMs are pre-trained on large natural language corpora, they have learned knowledge that can be transferred for other tasks and domains in zero-shot (i.e., no labeled data is available) or few shot settings (i.e., only a few labeled data is available for training).

Objective: In this study, we evaluate PLMs for issue classification from app reviews in multiple settings and compare them with the existing models.

Method: We set up different studies to evaluate the *performance* and *time efficiency* of PLMs compared to Prior approaches on six datasets. In addition, we train and study **domain-specific (Custom) PLMs** by incorporating app reviews in the pre-training. We report Micro and Macro Precision, Recall, and F1 scores and the time required for training and predicting with the models to answer three research questions.

RQ1: How accurate and efficient are the PLMs in classifying app reviews compared to the existing tools?

RQ2: How does the performance of the PLMs change when they are pre-trained on an app-review dataset instead of a generic dataset (e.g., Wiki documents, book corpus)?

RQ3: How do the PLMs perform in the following settings? Binary vs. multi-class setting, Zero-shot classification, Multi-task setting (i.e., different app-review analysis tasks), and Classification of user reviews collected from different resources (i.e., Twitter, App Store).

Results: Our results show that PLMs can classify the app issues with higher scores, except in multi-resource settings. On the largest dataset, results are improved by 13 and 8 micro- and macro-average F1-scores, respectively, compared to the Prior approaches. Domain-specific PLMs achieve the highest scores in all settings with *less* prediction time, and they benefit from pre-training with a larger number of app reviews. On the largest dataset, we obtained 98 and 92 micro- and macro-average F1-score (from 4.5 to 8.3 more F1-score compared to general pre-trained models), 71 F1-score in zero-shot setting, and 93 and 92 F1-score in multi-task and multi-resource settings, respectively, using the large domain-specific PLMs.

Conclusion: Although Prior approaches achieve high scores in some settings, PLMs are the only models that can work well in the zero-shot setting. When trained on the app review dataset, the Custom PLMs have higher performance and lower prediction times.

Index Terms: **Pre-trained Neural Language Models, App Review Classification.**

II. Presenter

Dr. Fatemeh Fard will present the work.

III. Original Paper

The DOI of the paper is 10.1007/s10664-023-10314-x, and here is the link to the paper https://doi.org/10.1007/s10664-023-10314-x.

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada RGPIN-2019-05175.