Two Datasets for Sentiment Analysis in Software Engineering

Bin Lin*, Fiorella Zampetti†, Rocco Oliveto‡, Massimiliano Di Penta†, Michele Lanza*, and Gabriele Bavota*
*Università della Svizzera italiana (USI), Switzerland — †University of Sannio, Italy — ‡University of Molise, Italy
*{firstname.lastname}@usi.ch — †fiorellazampetti@gmail.com, dipenta@unisannio.it — ‡rocco.oliveto@unimol.it

***Abstract*—Software engineering researchers have used sentiment analysis for various purposes, such as analyzing app reviews and detecting developers' emotions. However, most existing sentiment analysis tools do not achieve satisfactory performance when used in software-related contexts, and there are not many ready-to-use datasets in this domain. To facilitate the creation of better tools and sufficient validation of sentiment analysis techniques, we present two datasets with labeled sentiments, which are extracted from mobile app reviews and Stack Overflow discussions, respectively. The web app we created to support the labeling of the Stack Overflow dataset is also provided.**

I. Introduction

Recently, software engineering researches have adopted sentiment analysis techniques for various purposes, such as assessing the sentiment in app reviews, or analyzing developers' emotions. However, previous studies [1], [2] have disclosed that current sentiment analysis tools do not achieve satisfactory performance when used in software-related contexts. Therefore, customization and careful reliability verification are necessary when using existing sentiment analysis tools for software engineering studies. Nevertheless, there are currently not many ready-to-use datasets reporting the sentiment expressed in sentences extracted from software-related artifacts.

Given the strong need for datasets that can be used to (re-)train and evaluate sentiment analysis tools in a software engineering context, we present two new datasets and the web app we created to label the sentiment for the second dataset, which provides the possibility of further expanding the dataset. All the related artifacts can be downloaded from https://sentidata.github.io/. We also provide a longer version of this document better detailing the process used for construction of the datasets.

II. Datasets

A. Dataset of Mobile App Reviews

The first dataset contains 341 sentiment-annotated sentences from app reviews. The app reviews were originally collected by Villarroel *et al.* [3]. We manually labeled the sentiment of each review. Three scores are used to represent the sentiment: 1 for positive, 0 for neutral, and -1 for negative. The dataset contains 130 positive, 25 neutral, and 186 negative reviews.

B. Dataset of Stack Overflow Discussions

The second dataset contains 1,500 sentences extracted from Stack Overflow discussions and 20k intermediate nodes composing these sentences. All of them are sentiment-annotated.

We created this dataset to build a sentiment identification model based on `Stanford CoreNLP` [4] and customized for sentiment detection in Stack Overflow discussions. `Stanford CoreNLP` leverages a Recursive Neural Network (RNN), and was originally trained on the sentiment of movies' reviews. To train our new model, it is not sufficient to simply provide the polarity for a sentence, since the model needs to learn how sentences are grammatically built on top of positive/negative terms (*i.e.,* the polarity of all intermediate nodes composing a sentence is needed). An example can be found in Fig. 1. To use this sentence composed of 5 words, we had to label the sentiment of all 9 nodes depicted in Fig. 1. In the end, we labeled all 19,962 nodes composing these 1,500 sentences, with each node labeled by at least two people, with a third person resolving conflicts. In total, the dataset has 178 positive, 1,191 neutral, and 131 negative sentences. For the 1,500 sentences, three scores are used to represent the sentiment: 1 for positive, 0 for neutral, and -1 for negative. For the intermediate nodes, we present them in the Penn Tree Bank (PTB) format, which can be directly used by `Stanford CoreNLP`: 4 represents positive, 3 slightly positive, 2 neutral, 1 slightly negative, and 0 negative.

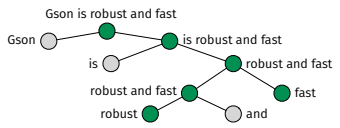

Fig. 1. Example of the labeling to build the Stanford CoreNLP training set.

We also provide the web app created for labeling the sentiment of nodes composing the sentences. The app will assign random nodes to app users, until each node is labeled by at least two people. Researchers can use this app to expand the dataset or customize their own labeling app.

References

[1] R. Jongeling, P. Sarkar, S. Datta, and A. Serebrenik, "On negative results when using sentiment analysis tools for software engineering research," *Empirical Software Engineering*, pp. 1–42, 2017.
[2] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, and R. Oliveto, "Sentiment analysis for software engineering: How far can we go?" in *Proceedings of ICSE 2018*, 2018, pp. 94–104.
[3] L. Villarroel, G. Bavota, B. Russo, R. Oliveto, and M. Di Penta, "Release planning of mobile apps based on user reviews," in *Proceedings of ICSE 2016*, 2016, pp. 14–24.
[4] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *In Proceedings of EMNLP 2013*, 2013.