Using Sentimental Analysis on Comments in Stack Overflow Posts to Identify if Post is Likely to be Edited in the Future

Eke, Norbert
NorbertEke@cmail.carleton.ca

Manes, Saraj Singh
sarajmanes@cmail.carleton.ca

February 25th, 2019

1 Introduction

Stack Overflow (SO) is the most popular question answering website for software developers, providing a large amount of code snippets and free-form text on a wide variety of topics. In the latest public data dump from December 9th 2018 SO listed over 42 million posts from almost 10 million registered users. Similar to other software artifacts such as source code files and documentation, text and code snippets on SO evolve over time. An example of this is when the SO community fixes bugs in code snippets, clarifies questions and answers, and updates documentation to match new API versions. To analyze how SO posts evolve Baltes et al. [2] built SOTorrent [1], an open data set aggregating and connecting the official SO data dump to other web sources, such as GitHub repositories. SOTorrent provides access to version histories of SO content at two separate levels: whole posts and individual text or code blocks.

Since the existence of SO in 2008, a total of 13.9 million SO posts have been edited after their creation. 19,708 of these posts have been edited even more than ten times. 300 million Software developers and engineers visit Stack Overflow monthly [3]. This number show the scale of interaction

happening at this platform. When a questions is asked or answered, most of discussion and interaction related to that topic happens in the form of comments. Comments can be associated with a question or an answer. These comments provide rich source of natural language text (mainly in English) to study developers' attitude towards a topic. As part of this research project we aim to explore this database to answer some research questions.

2 Motivation

At the Mining Software Repositories (MSR) conference in 2018, SOTorrent, a database built on entire SO content was released. In the submission paper [1] the authors performed top level analysis on this database. As part of [1], Baltes et al. claimed that out of all posts on SO, 38.6% have been edited after their creation. Furthermore, the authors of the paper argued that all edited posts are very rich in comments. They have large number of comments compared to non-edited posts. They inferred that these comments lead to the edit of the post. As part of this project we would like to perform very specific sentimental analysis on these text rich comments of edited post to argue about nature of these comments. In particular we are hoping to catch discontent or doubt in form of comment on SO post in question. We believe if exact sentiment of comment is identified, one can argue about reliability of information of presented in form of SO post.

3 Objective

As part of this project we would like to answer the following three research questions:
RQ1: For all edited posts on SO, what is the overall sentiment of comments around the editing time of the post? Does the sentiment change after the editing of the post?
RQ2: If the sentiment is negative, is there suspicious content in the post ? Is there something wrong the technique or concept within the post in question or is the code snippet wrong?
RQ3: Knowing if a post is reliable/stable (unlikely to be edited in future) is important for software engineers. How to identify/detect if a post will likely be edited in the future ?

RQ1 is closely related to general sentiment analysis on SO data. RQ2 looks more in depth at what could be wrong within the post. The real challenge with RQ2 is to design an approach using existing NLP techniques that can potentially detect anything suspicious activity within the post. Finally, for RQ3 does answering any of these research questions help to identify if a post is likely to be edited in the future ?

As a bonus task, if time will allow, we would like to take a look at similarity between word vectors of post answers and comments. Given an answer to a SO post and its comments we would like to explore if word vector similarities exist between core terms in the question. This would indicate the closeness of discussion to original topic.

References

[1] Baltes, S., Dumani, L., Treude, C., and Diehl, S. Sotorrent: reconstructing and analyzing the evolution of stack overflow posts. In *Proceedings of the 15th International Conference on Mining Software Repositories, MSR 2018, Gothenburg, Sweden, May 28-29, 2018* (2018), pp. 319–330.

[2] Baltes, S., Treude, C., and Diehl, S. Sotorrent: Studying the origin, evolution, and usage of stack overflow code snippets. *arXiv preprint arXiv:1809.02814* (2018).

[3] StackExchange. Stack exchange traffic statistics, 2018.