Building a Pilot Software Quality-in-Use Benchmark Dataset

Issa Atoum*, Chih How Bong, Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak
94300 Kota Samarahan,Sarawak,Malaysia
atoum@siswa.unimas.my, chbong@fit.unimas.my, nara@fit.unimas.my

***Abstract-* Prepared domain specific datasets plays an important role to supervised learning approaches. In this article a new sentence dataset for software quality-in-use is proposed. Three experts were chosen to annotate the data using a proposed annotation scheme. Then the data were reconciled in a (no match eliminate) process to reduce bias. The Kappa, statistics revealed an acceptable level of agreement; moderate to substantial agreement between the experts. The built data can be used to evaluate software quality-in-use models in sentiment analysis models. Moreover, the annotation scheme can be used to extend the current dataset.**

Keywords—Quality in use, Benchmark dataset, software quality, sentiment analysis

I. INTRODUCTION

Thrive on the World Wide Web and social media make Internet technology an invaluable source of business information. For instance, the product reviews on social media site composed collaboratively by many independent Internet reviewers through social media can help consumers make purchase decisions and enable enterprises to improve their business strategies. Various studies showed that online reviews have real economic values for targeted products .One type of reviews is the software reviews that covers users comments about used software.

Often users spend a lot of time reading software reviews trying to find the software that matches their needs (Quality-in-Use). With thousands of software published online it is essential for users to find quality software that matches their stated or implied requirements. Software Quality-in-Use (QinU) can be conceptually seen as the user point of view of software. It has gained its importance in e-government applications [1], mobile-based applications [2], [3], web applications [4], [5] and even business process development [6]. Prepared domain specific datasets plays an important role to supervised learning approaches.

Prepared dataset to this domain is essential to evaluate and coarse-grain results according to human perspectives. Literature has reported several datasets on diverse domains; movie reviews , customer reviews of electronic products like digital cameras [7] or net-book computers [8], services [9] , and restaurants [8], [10]. However, to the best of our knowledge, there are no datasets for software quality-in-use. Quality-in-use provides the viewpoint of the user on certain software. Moreover, our study to software review reveals that software reviews have several problems. Many of them are grammatically incorrect, they cover poor to rich semantic over different sentences, and they convey the user language that does not comply with the ISO standard definition of QinU[11]. To solve these problems an experiment was done using Google Search Engine (SE) to play the role of annotators by seeding the SE with keywords from the ISO Document. Conversely, results were poor and that was the main motive for preparing a dataset to be used in supervised learning mode.

This work proposes a new gold standard dataset for software quality-in-use built through an annotation scheme. The gold standard dataset here is a set of software reviews crawled from the web and classified by human experts (annotators). The objective of this dataset is to be able to compare the results of the proposed method versus the data that is manually annotated by experts. The building process starts with software reviews and ends up with labeled sentences. At the end of the annotation process, each software review-sentence will have the sentence QinU *topic* (*characteristic*), sentence polarity (*positive, negative* or *neutral*) and indicating topic *features*. First, a set of reviews from different categories are crawled from Amazon.com and Cnet.com respectively to cover the software reviews domains. These reviews are filtered from junk and non-English text. Next, a balanced set of reviews per rate is selected. Then, reviews are split into sentences. Finally, the sentences are classified by annotators and sentence classification data is saved in the Database.

First Related works are summarized. Next software reviews and annotators selection processes are