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
The pros, cons and summary serve a place for positive, negative and summary of Cnet reviews.

2 Natural Language Toolkit (nltk version 2.0.4) under python 2.7.8. http://www.nltk.org
topic efficiency. Additionally the annotator will choose why a sentence was positive or negative by choosing an opinion word (Step4). Finally the classified sentences are saved in the database for the next step which is QinU identification (Step5).

VI. DATA RECONCILIATION
In this dataset, the QinU characteristics; effectiveness, efficiency and freedom from risk are taken while satisfaction and context coverage are not considered. The reasons are 1) It is found that satisfaction can be due to additional services aside from the product like the software price, the delivery, download website or just a word of the mouth (“it is good”, but why no why!!), and 2) It is assumed that when the software is filling the three characteristic then it is supposed to fill satisfaction as well according to the ISO standard definition. The satisfaction is deferred to be undertaken in future work.

The Context coverage is not taken into consideration because it is assumed that all users have the same level of understanding and they have followed the software installation guides before they place their reviews.

Given the initial training dataset, it is found that it can happen that the annotators cannot agree on the sentence topic, feature(s), and sentiment orientation, so a reconciliation process is desirable. A manual reconciliation is labor costly and can be biased due to human nature. Moreover, final agreement might not be achievable due to sentences that are congenital difficult (rich semantic). Therefore, the sentences follow the steps shown in Fig. 2 before they are used as a final gold standard.

In Fig. 2 a set of three sentences is taken from the three annotators (Ann1, Ann2, and Ann3). Then the merge process starts by first choosing the majority topic between sentences. Then features named (F11, F12, F13) that were chosen by the first annotator (Ann1) are merged with features from the second and third annotators (F21, F22, F23; F31, F32, F33) respectively. The common features are only chosen. The same is done for the sentence polarity (sentiment), Polarity1, Polarity2 and Polarity 3. Finally the gold standard knowledge base will contain sentence majority topic, common features and majority polarity.

Step1: Merge the three annotator’s sentences topics by choosing the majority topic. Sentences that have no identical topic are eliminated from the dataset because it might be very context sensitive or it can give different understanding between annotators

![Fig. 2 Data Reconciliation](image-url)
**Step2:** From the merged sentences (with majority topic), choose the list of common features. If no common features are found then the merged sentence is disused.

**Step3:** From the merged sentences (with majority topic and common features), set the merged sentence polarity to majority polarity. Positive and negative sentences become neutral.

### VII. EXPERTS AGREEMENT

To justify the validity of the resultant data set, the Kappa (\(\kappa\)) agreement[17] is used. The Kappa measures the degree of agreement known as (inter-annotator agreement) between annotators [18]. The Cohen Kappa for topics agreement are calculated and it was found as shown in Table I. According to the guidelines of [19], the agreement is between moderate to substantial agreement. Hence, the agreement is acceptable given three different topics and three experts.

| Group                  | Kappa, Interpretation |
|------------------------|-----------------------|
| Annotator1,Annotator2  | 0.46 Moderate         |
| Annotator1,Annotator3  | 0.58 Moderate         |
| Annotator2,Annotator3  | 0.69 Substantial      |

### VIII. SAMPLE DATASET DESCRIPTION

**APPENDIX I**

Sample of quality in use dataset*.

| ID# | sentence                                                                 | topic | Polarity |
|-----|--------------------------------------------------------------------------|-------|----------|
| 1   | Upgrade because 2010 [worked] well, so other than col, no need to upgrade.| 1     | -1       |
| 2   | making receipts [optional].                                              | 1     | 1        |
| 3   | in addition, many of the [features] are so automated as to be useless   | 1     | -1       |
|     | or worse.                                                                |       |          |
| 4   | Aside from the mobile app not [working]                                  | 1     | -1       |
| 5   | i am using citibank's financial [tool] instead.                          | 1     | -1       |
| 6   | not much [different] from quicken 2010!                                  | 1     | -1       |
| 7   | must search for [functions] that were [prior] on the screen.             | 1     | 0        |
| 8   | i think the [layout] was better in the older version.                    | 1     | 0        |
| 9   | it's more [difficult] to find functions i used to frequently need.      | 1     | -1       |
| 10  | don't [need] all the other detailed [items].                             | 1     | -1       |
| 11  | if [performs] as expected.                                               | 1     | 1        |
| 12  | i like the new [look]                                                   | 1     | 1        |
| 13  | the [color] [schemes] are absolutely atrocious!                          | 1     | -1       |
| 14  | you have your [choice] of 3 [color] [options]: white, gray and light gray.| 1     | 0        |
| 15  | the white is so blindingly white it's [hard] on the eyes.                | 1     | -1       |

* Features in brackets[]:Sentences classified as effectiveness(1), efficiency(2), risk(3) (column 3). Polarity -1:negative, 1 positive and 0 neutral

Table II shows a sample of sentences and their QinU topics (features are in brackets []). The complete dataset will be available for download at this [http://www.meta-net.eu/](http://www.meta-net.eu/). The dataset have reviews, sentences, topics, features, polarities, modifiers and other data as well. Appendix I shows the pilot dataset.

| Sentence                                      | Topic          |
|-----------------------------------------------|----------------|
| the [color] [schemes] are absolutely atrocious!| Effectiveness  |
| OpenOffice is [fast]                         | Efficiency     |
| [crash] too often especially when opening ms  | Risk           |
| office files.                                |                |

### IX. CONCLUSION

In this article, the gold standard building process is illustrated. Three experts were chosen to annotate the data through a proposed annotation scheme. 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 could be used to evaluate software quality models. It is found that some software reviews could have high sentence semantics resulted from different profile users and sentence interconnections. Thus, further research is needed to face the challenges of satisfaction, and context coverage characteristics and further enhance sentence splitters.
| ID# | sentence                                                                                                                                                                                                 | topic | Polarity |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----------|
| 16  | it's [hard] to know exactly where you are in any of the programs.                                                                                                                                           | 1     | -1       |
| 17  | the [look] of this was so shocking                                                                                                                                                                         | 1     | -1       |
| 18  | ms touts the cloud [options] but to most users that's not really going to benefit them.                                                                                                                      | 1     | -1       |
| 19  | what's the point of having updating tiles if they don't [work] with your email program.                                                                                                                    | 1     | -1       |
| 20  | so now i am kinda p'o'd but other than that is [sparkles] for office.                                                                                                                                       | 1     | 0        |
| 21  | it is [slower] too.                                                                                                                                                                                          | 2     | -1       |
| 22  | also all the programs seem to run [slower] than 2010.                                                                                                                                                      | 2     | -1       |
| 23  | office 2013 doesn't [integrate] with the windows metro tiles.                                                                                                                                               | 2     | -1       |
| 24  | is taking more [time] than i'd like to adjust to it.                                                                                                                                                       | 2     | -1       |
| 25  | but it's a way [faster] than 2010                                                                                                                                                                          | 2     | 1        |
| 26  | [quick] install and [startup] easy.                                                                                                                                                                          | 2     | 1        |
| 27  | the first problem is that access 2010 is about three times [slower] than access 2002 when operating in win 7 pro.                                                                                           | 2     | -1       |
| 28  | but version 12 is much [faster]                                                                                                                                                                            | 2     | -1       |
| 29  | this product needs some significant [time] and patience to get up and [running].                                                                                                                          | 2     | -1       |
| 30  | bought dragon for a friend and found out his computer did was not [compatible].                                                                                                                             | 2     | -1       |
| 31  | the game ran very [smooth].                                                                                                                                                                                 | 2     | 1        |
| 32  | it's called [load] and [performance] testing.                                                                                                                                                               | 2     | 0        |
| 33  | instead it has a more [time] [consuming] layer [process].                                                                                                                                                   | 2     | -1       |
| 34  | its [slowness] is maddening.                                                                                                                                                                                | 2     | -1       |
| 35  | the “transformation” into a [fast] pouncing snow leopard gave me an [additional] 11[gb] [space] on my [hard drive].                                                                                           | 2     | 1        |
| 36  | so there was a 9[gb] [increase] in my available [hard drive] [space].                                                                                                                                       | 2     | 0        |
| 37  | i thought that maybe this game wasn't [supported] by 10.6.6.                                                                                                                                               | 2     | -1       |
| 38  | it probably would [slow down] imovie and final cut as well.                                                                                                                                                 | 2     | -1       |
| 39  | i'm currently [running] a 2007 24 inch imac with the intel 2.33[ghz] core 2 duo, 3[gb] [ram] and nvidia geforce 7600 gt.                                                                                 | 2     | 1        |
| 40  | the majority of improvements affect system [reliability], [speed], and [resourcefulness].                                                                                                                   | 2     | 0        |
| 41  | while quicken seems to have resolved the upgrade [issues]                                                                                                                                                   | 3     | 1        |
| 42  | it has some minor [bugs] , but nothing that can't be overcome.                                                                                                                                              | 3     | 0        |
| 43  | the program is good, but quicken clearly has some serious [issues] with [connectivity] both with their servers and the program itself.                                                                | 3     | -1       |
| 44  | it's just too [buggy].                                                                                                                                                                                      | 3     | -1       |
| 45  | no [issues] with install                                                                                                                                                                                   | 3     | -1       |
| 46  | i do not want to be [forced] to have pop-up [warnings].                                                                                                                                                     | 3     | -1       |
| 47  | it has a few tiny [glitches] - not really glitches but it, of course,                                                                                                                                        | 3     | 0        |
| 48  | this is most likely flash's [fault].                                                                                                                                                                          | 3     | -1       |
| 49  | the last two versions (12 and 11)cause a [hanging] problem that dragon has been incapable of fixing.                                                                                                         | 3     | -1       |
| 50  | then [hangs] / [stalls].                                                                                                                                                                                   | 3     | -1       |
| 51  | who in their right [mind] stores important program files in a [temp] directory?                                                                                                                              | 3     | -1       |
| 52  | this program [hogs] a lot of memory                                                                                                                                                                          | 3     | -1       |
| 53  | ea should have some kind of [backup] plan that would allow the game to be played offline while the servers are [fixed].                                                                                      | 3     | 0        |
| 54  | i have not had any server [issues] like other reviews have stated.                                                                                                                                         | 3     | 1        |
| 55  | have never had a server [issue].                                                                                                                                                                            | 3     | 1        |
| 56  | i hope they [fix] it soon                                                                                                                                                                                  | 3     | 1        |
| 57  | some game mechanics appear to be [broken]; i. e. traffic, r vs. c vs. i zoning (you can make completely residential cities), trading, emergency services, etc.                                                  | 3     | -1       |
| 58  | see the official forums for a list of [bugs].                                                                                                                                                                | 3     | 0        |
| 59  | [excessive] drm doesn't stop [pirates] it makes them.                                                                                                                                                      | 3     | -1       |
| 60  | buy at your own [risk].                                                                                                                                                                                    | 3     | -1       |
REFERENCES

[1] N. B. Osman and I. M. Osman, “Attributes for the quality in use of mobile government systems,” in 2013 International Conference on Computing, Electrical and Electronics Engineering (ICCEE), 2013, pp. 274–279.

[2] R. Alnanih, O. Ormandjieva, and T. Radhakrishnan, “A New Methodology (CON-INFO) for Context-Based Development of a Mobile User Interface in Healthcare Applications,” in Pervasive Health SE - 13, 1st ed., A. Holzinger, M. Zeile, and C. Röcker, Eds. London: Springer London, 2014, pp. 317–342.

[3] H. J. La and S. D. Kim, “A model of quality-in-use for service-based mobile ecosystem,” in 2013 1st International Workshop on the Engineering of Mobile-Enabled Systems (MOBS), 2013, pp. 13–18.

[4] T. Oreho, A. Granic, and D. Kermek, “Evaluating the perceived and estimated quality in use of Web 2.0 applications,” J. Syst. Softw., vol. 86, no. 12, pp. 3039–3059, Dec. 2013.

[5] C.-C. Hsu and Y.-S. Lee, “Exploring the Critical Factors Influencing the Quality of Blog Interfaces Using the Decision-Making Trial and Evaluation Laboratory (DEMATEL) Method,” Behav. Inf. Technol., vol. 33, no. 2, pp. 184–194, 2014.

[6] R. Heinrich, A. Kappe, and B. Pach, “Modeling Quality Information within Business Process Models,” in Proceedings of the 4th SQMB Workshop, TUM-I1104, 2011, pp. 4–13.

[7] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[8] S. Brody and N. Elhadad, “An unsupervised aspect-sentiment model for online reviews,” in Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 804–812.

[9] C. Long, J. Zhang, and X. Zhut, “A Review Selection Approach for Accurate Feature Rating Estimation,” in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 766–774.

[10] G. Ganu, N. Elhadad, and A. Marian, “Beyond the stars: Improving rating predictions using review text content,” in 12th International Workshop on the Web and Databases, 2009.

[11] ISO/IEC, “ISO/IEC 25010: 2011, Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models,” International Organization for Standardization, Geneva, Switzerland, 2011.

[12] I. Atoum, C. H. Bong, and N. Kulathuramaiyer, “Towards Solving Software Quality-in-Use Measurement Challenges,” J. Emerg. Trends Comput. Inf. Sci., vol. 5, no. 11, pp. 877–885, 2014.

[13] I. Atoum and C. H. Bong, “Measuring Software Quality in Use: State-of-the-Art and Research Challenges,” ASQ Software Qual. Prof., vol. 17, no. 2, pp. 4–15, 2015.

[14] T. T. Thet, J.-C. Na, and C. S. Khoo, “Aspect-based sentiment analysis of movie reviews on discussion boards,” J. Inf. Sci., vol. 36, no. 6, pp. 823–848, 2010.

[15] I. Atoum and C. H. Bong, “A Framework to Predict Software ‘Quality in Use’ from Software Reviews,” in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), vol. 285, J. Herawan, Tutu and Deris, Mustafa Mat and Abawajy, Ed. Kuala Lumpur: Springer Singapore, 2014, pp. 429–436.

[16] S. T. W. Wendy, B. C. How, and I. Atoum, “Using Latent Semantic Analysis to Identify Quality in Use (QU) Indicators from User Reviews,” in The International Conference on Artificial Intelligence and Pattern Recognition (AIIPR2014), 2014, pp. 143–151.

[17] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.

[18] N. C. Smeeton, “Early history of the kappa statistic,” Biometrics, vol. 41, no. 795, 1985.

[19] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, no. 1, p. 159, 1977.