This is a post-peer-review, pre-copyedit version of this article. The final authenticated version is available online at:
https://doi.org/10.1109/REW53955.2021.00052

The Use of Sub-forums in Software Product Forums

Hechen Wang, Peter Devine, James Tizard, Seyed Reza Shahamiri, Kelly Blincoe
Human Aspects of Software Engineering Lab
University of Auckland, New Zealand
{hwan531, pdev438, jtiz003}@aucklanduni.ac.nz, admin@rezanet.com, k.blincoe@auckland.ac.nz

***Abstract*—Software product forums is a platform filled with user feedback that utilises the sub-forum feature to categorise user discussion into themes. These sub-forums are very similar to classification labels that have been used to automatically classify user feedback on other platforms such as Troubleshooting and Feature Request. It would be very beneficial to the CrowdRE community if these sub-forum categories can be utilised in a research setting as it would reduce the effort required to label content for classification manually. However, no research has been done on the accuracy of these sub-forum categorisations in software product forums. In this exploratory study, we examined the accuracy of user categorised posts in two software product forums and discovered that users incorrectly categorise more than 20% of the posts during submission. Our discovery suggests that at the current stage, sub-forum categories should not be trusted as a label to classify feedback automatically.**

***Index Terms*—software product forum, user feedback, Requirements Engineering, CrowdRE**

I. Introduction

Reviews and discussions around software products often contain valuable information for software developers. While it is possible to examine all discussions by hand, such time-intensive tasks are often better suited for automation so developers can use their time more effectively[1]. Recent research has mostly focused on examining content from app stores and Twitter to automatically extract product development insights (e.g.[2],[3]). These approaches have primarily relied on classification techniques to automatically group related feedback into pre-defined categories (like bugs or feature requests)[4]. Software product forums, a type of online question and answer forum where software users can discuss specific software products, are another platform that contains product development insights. These forums have not been studied as much despite many developers actively using software product forums to communicate with their users[5, 6, 7]. Compared to other platforms, most software product forums contain sub-forums to divide discussions into themes to allow better grouping of topics[8]. Many of these sub-forum categories are similar to the high-level classification labels that have been used to automatically classify feedback on other platforms, such as *"Troubleshooting"* and *"Feature Request"*[9]. Compared to prior studies, where models that automatically classify feedback have been trained and tested using datasets that were labelled through extensive manual content analysis, the sub-forum categories in software forums could potentially be used with no such manual effort. Thus, these labels have promise for use in CrowdRE research. However, such labels can only be used to automatically group related feedback if users are correctly putting their forum posts in the appropriate sub-forums. Since forums currently have no way to enforce the correct selection of appropriate sub-forums, users could potentially submit their user feedback in the wrong sub-forum. For example, users could submit a feature request in the Troubleshooting sub-forum by mistake. Since prior research has not yet examined the accuracy of the grouping of forum posts into sub-forums, it is not clear if these sub-forums can be used to help group related forum posts automatically.

To understand the potential of sub-forums as an initial high-level classification of forum posts, we perform an exploratory study on the use of sub-forums and the correctness of the posts submitted by users in sub-forums on two software product forums. We examined the number of incorrect sub-forum post submissions within the VLC Media Player forum[1] and Spotify Help forum[2] by sampling a small number of post titles. We found that 22% and 27.8% of forum posts within those two forums are placed in the incorrect sub-forum by users during submission. We further discuss how awareness of potentially miscategorized forum posts is important for CrowdRE researchers.

II. Related Work

User feedback can be seen as a form of communication between developers and software users. Successful communications are essential in software product development [10]. However, incorrect information provided by users can hinder the development process, as failure to communicate effectively is often cited as one of the major reasons for software development failures [11]. In other fields, it has been found that up to 30% of communications can be categorised as communication failures, causing stress and further issues in procedures [12]. For software development, Chari and Agrawal's work examined the impact of incorrect requirements on waterfall software project outcomes and discovered that incorrect software requirements increase the number of new requirements as well as the number of defects injected [13]. Incorrect categorisation of software product forum threads can present incorrect information to developers that hinder software development processes.

Within the field of Requirements Engineering, user feedback is often used to extract valuable software insight to

[1] https://forum.videolan.org/viewforum.php?f=21
[2] https://community.spotify.com/t5/Help/ct-p/Help

developers [7]. Due to the amount of user feedback that is available on platforms such as Google Play Store and Twitter, the task of understanding reviews can no longer be achieved through manual review analysis, raising the need for software user feedback classification to group feedback into software requirement related labels for faster processing [14]. Studies often use platform-specific features, such as ratings and likes from Play Store and Twitter, to help with the classification task [4]. For software product forums, sub-forum categories can be seen as a key feature of the platform, yet to our understanding, no research has studied the accuracy of these categorisations by users. Examining the use of these sub-forum categories by users in software forums allows us to better understand the platform and present pathways for future research in this area.

III. Data Collection

For this study, we chose two popular software product forums to collect our data, the VLC Media Player forum and Spotify Help Forum. These forums were selected since they both cover a wide range of topics from many users over an extended period of time. Both forums have been active for more than five years with more than 100,000 registered users on each forum, and they have a wide range of sub-forums. From the VLC Media Player forum, we collected all of post titles from two of its sub-forum categories, *Windows Troubleshooting* and *Feature Request* on 03/06/2020. For Spotify Help forum, we collected all of the post titles from six of its sub-forum categories, namely *Accounts*, *Subscriptions*, *Premium Family*, *Premium Student*, *Windows Troubleshooting*, and *Ideas(Feature Request)* on 06/06/2020. Troubleshooting and Feature Request sub-forums were chosen for both forums since they closely resemble high-level classification labels (bug and feature requests) found in similar studies [9]. In addition to those two sub-forums, we also referred to the FAQs for both VLC[3] and Spotify[4] to examine commonly discussed issues for each software system and discovered that account and subscription issues are heavily discussed on the Spotify forum. Therefore, in addition to the Troubleshooting and Feature Request sub-forums, we also collected post titles from Accounts, Subscriptions, Premium Family, and Premium Student sub-forums to ensure that these popular issues are included in our dataset. In total, we collected 48,053 and 114,887 thread titles from VLC and Spotify Help forums.

IV. Research Methodology

From our collected data, we randomly sampled a small subset for manual labelling as showing in Table I. The size of our sample set was determined by calculating the population needed to reach at least a 95% confidence level with a confidence interval of 10% [15]. To ensure a balanced sample set where each sub-forum category has the same chance to be sampled, we randomly sampled 2,000 posts from each of the sub-forums collected, resulting in 4,000 posts for VLC Media Player forum and 12,000 posts for the Spotify Help forum. We then randomly sampled 400 posts for each forum from this dataset for manual analysis to meet the confidence requirement. In total, 800 post titles were manually analysed for this study.

TABLE I
Overview of Data Collection

Forum	Sub-Forum Category	#Titles	#Trimmed	Sample
VLC	Windows Troubleshooting	42021	2000	176
	Feature Request	6032	2000	234
Total		**48053**	**4000**	**400**
Spotify	Accounts	70161	2000	66
	Subscriptions	19153	2000	60
	Premium Family	7800	2000	48
	Premium Student	3641	2000	68
	Windows Troubleshooting	11796	2000	75
	Ideas	2336	2000	83
Total		**114887**	**12000**	**400**

For the creation of the truth set, we conducted Manual Content Analysis to examine the content of each post title within our sample [16]. Two PhD students with deep knowledge in user feedback independently analysed and classified each post title for both the VLC and Spotify data samples. Each coder was presented with a list of categories and a set of post titles. They were asked to examine each post title carefully and then classify the title into a category. For VLC samples, the categories were *Windows Troubleshooting* and *Feature Request*. For Spotify samples, the categories were *Accounts*, *Subscriptions*, *Premium Family*, *Premium Student*, *Windows Troubleshooting*, and *Feature Request*. The categories were chosen to match the sub-forums from which the data were collected, allowing us to examine the accuracy of these sub-forum categories. This process was performed in three rounds, starting with coding 10% of the post titles together to establish a baseline between coders. Then each coder individually coded the rest of the dataset before meeting up to discuss initial disagreement between coders. Each coder then individually updated their labels before having a final discussion to reconcile disagreements. The intercoder reliability was calculated using Cohen's Kappa since the data was nominal [17]. The ReCal2 tool[5] was used to calculate the Kappa scores [18]. As shown in Table II, there was a strong level of agreement between coders for both datasets since the Kappa values are over 0.8 [19]. When the two coders could not reach an agreement after the final discussion or both coder agreed that the title did not belong to any of the given sub-forum categories, the post was discarded from our truth set. In total, 31 and 12 post titles were discarded from the VLC and Spotify sample set, with 369 and 388 post titles chosen as our truth set for this study. The main reason for disagreements between coders was the ambiguity of the post title, allowing it to fit into multiple sub-forum categories. We then compared our truth set classification of each post title to its originating sub-forum category.

[3] https://www.videolan.org/support/faq.html1
[4] https://community.spotify.com/t5/FAQs/tkb-p/Spotify-Answers
[5] http://dfreelon.org/utils/recalfront/recal2/

TABLE II
INTERCODER RELIABILITY

Data set	Initial Agreement	Reconciled Agreement	Cohen's Kappa
VLC	70.9%	92.5%	84.8%
Spotify	84%	97%	96.4%

V. RESULTS

Table III presents the difference between user categorised posts and our truth set classification. We discovered that 22.0% (81) and 27.8% (108) of post titles within our sample set of VLC and Spotify forums were incorrectly categorised by users during post submission. Figure 1 shows an example of an incorrectly submitted post by a user on the VLC forum, where a bug report was posted in the feature request forum. Figure 2 presents an example where the user is unaware of an existing feature, therefore submitting a feature request. Figures 3 and 4 present the confusion matrices between the user categorised posts and the truth set classifications for the VLC and Spotify forums. For the VLC Media Player forum, users have posted a good amount of troubleshooting posts on the feature requests sub-forum. For the Spotify forum, account based issues are being submitted to the Premium Family and Premium Student sub-forums.

TABLE III
TRUTH SET VS SUB-FORUM CATEGORY

Forum	Sub-Forum Category	#User Categorised	#Truth Set
VLC	Windows Troubleshooting	210	263
	Feature Request	159	106
Spotify	Accounts	62	32
	Subscriptions	59	79
	Premium Family	46	50
	Premium Student	68	52
	Windows Troubleshooting	71	97
	Feature Request	82	78

Fig. 1. Troubleshooting post submitted as feature request [6]

Fig. 2. Incorrectly categorised post in VLC forum [7]

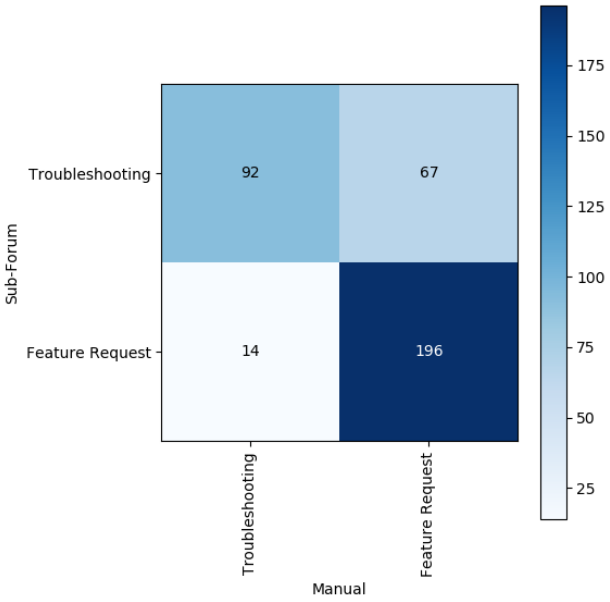

Fig. 3. Confusion Matrix for VLC user submissions

VI. DISCUSSION

In this section, we describe some of the findings and their implications that emerged from this study.

1) Sub-forum categories are not always reliable: With the amount of user feedback available online, recent research has moved towards using automation to better extract software requirements from the large amount of user feedback. Methods for extracting requirements from App Stores and Twitter have found adding meta-data, such as app ratings and text

[6] https://forum.videolan.org/viewtopic.php?f=7&t=30185
[7] https://forum.videolan.org/viewtopic.php?f=7&t=102139

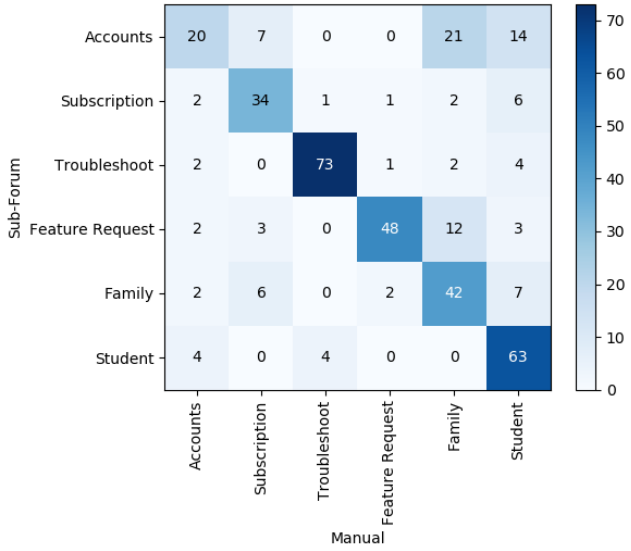

Fig. 4. Confusion Matrix for Spotify user submissions

length, improves the accuracy[20]. Sub-forum categories could be a promising type of meta-data to improve extraction of requirements from online product forums. On the forums we examined, the sub-forum topics share many similarities with high-level classification labels that are often found in related studies. Labels such as *Bug Report* and *Feature Request* are almost identical to the sub-forum categories of *Troubleshooting* and *Feature Request*. Thus, if the sub-forum categories are accurate, they could serve as a high-level classification of requirements without the need for manual labelling and training. However, we found that the sub-forum classifications of posts in software product forums are not always accurate. With over 20% of all posts being submitted in the wrong sub-forum, any attempts at high-level feedback classification using these categories will require more effort from researchers to ensure that each post selected is correctly categorised. Future CrowdRE research on requirement extraction from online product forums should take care in using sub-forum categories as a feature in any automatic classification attempts. Based on our findings, more research can be done to understand why users are submitting forum posts to incorrect sub-forums. This understanding could also help to devise mitigation strategies to ensure higher accuracy of sub-forum classifications in future posts.

2) Usability of forums: About one in every four posts submitted in software forums was submitted in the incorrect sub-forum. This suggests that forums should be improved to make them easier to use.

Recommendation 1: Ensure sub-forum topics are distinct and well described. For example, Spotify's subscription services are based around Premium monthly plans[8], which offers the choice of Premium, Premium Student, or Premium Family. On the forum, there are sub-forums for Subscriptions, Premium Student, and Premium Family. Since Premium Student and Premium Family are two types of subscriptions offered, there seems to be some confusion on where issues with these types of subscriptions should be posted. It is unclear if issues with a Premium Student subscription, for example, should be posted on the Subscription sub-forum or the Premium Student sub-forum. Ensuring detailed descriptions on what should be posted on each sub-forum as well as ensuring clear boundaries across each sub-forum can better support users in selecting the correct sub-forum for their post.

Recommendation 2: Use automation to support users when writing new posts. In both forums, we found that it is very common for senior community members to assist other members. Often, these members will respond to forums posts pointing to similar discussions when users post similar topics as previous posts. Adding features such as automatic duplication detection or sub-forum selection can make the forum much easier to use. Such features would enable users to join the discussion on existing posts or find answers to their questions right away, instead of posting duplicate issues which can flood the forum space. Of course, some requirement prioritisation techniques may currently utilise the number of posts about a particular issue as a way of understanding the scale of the issue. We recommend that if adding duplicate post detection, additional interactive features also be implemented such as a like button or ways for users to indicate that "I have the same issue". This would allow developers to still understand the scale of the issue.

Benefits to CrowdRE community: In the context of the CrowdRE community, making software product forums easier to use will enable product improvements to be extracted more efficiently as more posts would be in the correct sub-forum. Less duplicate topics of discussion would also mean that similar information would be grouped together instead of spread into different posts.

A. Threats to Validity

The main threat to validity for this study is that the findings are derived from only two unique forums. We cannot claim that the results generalise to other software product forums. It is possible that other software product forums have different structures and moderation levels and that the results of this study will not generalise to other online product forums. We chose our two forums from different fields with different forum structures, but future work can validate whether incorrect sub-forum categories are used on additional forums.

Another threat of this research is that we only examined the post title from each post to manually classify it into the categories. It is possible that the content within the post is different from the post title, which would affect the results of

[8] https://support.spotify.com/nz/account_payment_help/subscription_options/

this study. We tried to mitigate this limitation by discarding thread titles that were ambiguous and caused disagreements between the coders. Future work can perform more detailed analysis of forum posts to validate our findings.

VII. Conclusion

In this exploratory study on the use of sub-forums in software product forums, we examined the accuracy of user categorised posts within the VLC and Spotify Help forums. Through manual content analysis of 800 forum posts from different sub-forum categories, we discovered that users incorrectly categorise 22.0% and 27.8% of posts during submission. The insights we present from this study serve as a warning for future work on software product forums for CrowdRE researchers. We suggest more work can be done to improve software forums to reduce the number of incorrect categorisations in sub-forum categories. At this stage, researchers using sub-forum categories as high-level classification labels need to be aware of the number of incorrectly submitted posts by users.

References

[1] E. C. Groen, N. Seyff, R. Ali, F. Dalpiaz, J. Doerr, E. Guzman, M. Hosseini, J. Marco, M. Oriol, A. Perini *et al.*, "The crowd in requirements engineering: The landscape and challenges," *IEEE software*, vol. 34, no. 2, pp. 44–52, 2017.

[2] E. Guzman, R. Alkadhi, and N. Seyff, "An exploratory study of twitter messages about software applications," *Requirements Engineering*, vol. 22, no. 3, pp. 387–412, 2017.

[3] E. Guzman, R. Alkadhi, and N. Seyff, "A needle in a haystack: What do twitter users say about software?" in *2016 IEEE 24th International Requirements Engineering Conference (RE)*, 2016, pp. 96–105.

[4] C. Wang, M. Daneva, M. van Sinderen, and P. Liang, "A systematic mapping study on crowdsourced requirements engineering using user feedback," *Journal of software: Evolution and Process*, vol. 31, no. 10, p. e2199, 2019.

[5] J. Tizard, H. Wang, L. Yohannes, and K. Blincoe, "Can a conversation paint a picture? mining requirements in software forums," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019, pp. 17–27.

[6] S. Gottipati, D. Lo, and J. Jiang, "Finding relevant answers in software forums," in *2011 26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011)*. IEEE, 2011, pp. 323–332.

[7] J. Tizard, T. Rietz, and K. Blincoe, "Voice of the users: A demographic study of software feedback behaviour," in *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 2020, pp. 55–65.

[8] P. Holtz, N. Kronberger, and W. Wagner, "Analyzing internet forums," *Journal of Media Psychology*, 2012.

[9] D. Pagano and W. Maalej, "User feedback in the appstore: An empirical study," in *2013 21st IEEE International Requirements Engineering Conference (RE)*, 2013, pp. 125–134.

[10] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally distributed software development," *IEEE Transactions on software engineering*, vol. 29, no. 6, pp. 481–494, 2003.

[11] M. Fabriek, M. v. d. Brand, S. Brinkkemper, F. Harmsen, and R. Helms, "Reasons for success and failure in offshore software development projects," 2008.

[12] L. Lingard, S. Espin, S. Whyte, G. Regehr, G. R. Baker, R. Reznick, J. Bohnen, B. Orser, D. Doran, and E. Grober, "Communication failures in the operating room: an observational classification of recurrent types and effects," *BMJ Quality & Safety*, vol. 13, no. 5, pp. 330–334, 2004.

[13] K. Chari and M. Agrawal, "Impact of incorrect and new requirements on waterfall software project outcomes," *Empirical Software Engineering*, vol. 23, no. 1, pp. 165–185, 2018.

[14] C. Stanik, M. Haering, and W. Maalej, "Classifying multilingual user feedback using traditional machine learning and deep learning," in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2019, pp. 220–226.

[15] R. V. Krejcie and D. W. Morgan, "Determining sample size for research activities," *Educational and psychological measurement*, vol. 30, no. 3, pp. 607–610, 1970.

[16] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[18] D. G. Freelon, "Recal: Intercoder reliability calculation as a web service," *International Journal of Internet Science*, vol. 5, no. 1, pp. 20–33, 2010.

[19] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[20] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? on automatically classifying app reviews," in *2015 IEEE 23rd international requirements engineering conference (RE)*. IEEE, 2015, pp. 116–125.