A Taxonomy for User Feedback Classifications

Rubens Santos, Eduard C. Groen, Karina Villela
Fraunhofer IESE, Kaiserslautern, Germany
rubens.santos@iese-extern.fraunhofer.de;
{eduard.groen; karina.villela}@iese.fraunhofer.de

Abstract

Online user feedback contains information that is of interest to requirements engineering (RE). Natural language processing (NLP) techniques, especially classification algorithms, are a popular way of automatically classifying requirements-relevant contents. Research into this use of NLP in RE has sought to answer different research questions, often causing their classifications to be incompatible. Identifying and structuring these classifications is therefore urgently needed. We present a preliminary taxonomy that we constructed based on the findings from a systematic literature review, which places 78 classifications categories for user feedback into four groups: Sentiment, Intention, User Experience, and Topic. The taxonomy reveals the purposes for which user feedback is analyzed in RE, provides an initial harmonization of the vocabulary in this research area, and may inspire researchers to investigate classifications they had previously not considered. This paper intends to foster discussions among NLP experts and to identify further improvements to the taxonomy.

1 Introduction

Recent years have seen a growing interest in the requirements engineering (RE) community in online user feedback as a source of user requirements regarding software products, which is being studied specifically in the context of Crowd-based RE (CrowdRE) [GSA+17]. Research in this area has provided evidence that user feedback on a product and/or its competitor products contains sufficient RE-related information, irrespective of whether this feedback is obtained from (app) review sites [GKH+17, IH13, JM17], social media platforms such as Twitter [WM17, GIG17], or (bug) tracking systems [MFH+16, WZL+18]. Although the large quantities in which user feedback is available can be seen as a benefit, they also warrant the use of automated natural language processing (NLP) techniques rather than manual inspection [GSK+18]. Especially *classification algorithms* that categorize texts into predetermined categories seem to be suited for this purpose [JM17, LL17]. For RE, such categories could include "Bug Report", "Feature Request", and "Praise" [MN15].

Classification algorithms and other NLP techniques were not originally designed to support user feedback analysis for RE. This is why much of the emergent research in this field is dedicated to tailoring them to this purpose. A common denominator among these research efforts is that they seek to separate *requirements-relevant* content from *requirements-irrelevant* material [CLH+14]. The share of requirements-relevant content in user feedback is typically about 20–30% (e.g., [GKH+17, GIG17]), meaning that such a distinction already helps to greatly reduce the amount of user feedback to only such content that is relevant for RE purposes (e.g.,

summarization; [Ber17]). However, nearly all research has sought to provide more detailed sub-classifications of requirements-relevant content (see Section 3 for a review). It is at the level of these predefined categories, or *classifications*, that CrowdRE research diverges, by using classifications that are at best only modestly compatible with those of other works analyzing user feedback from different RE-relevant perspectives. However, these differences make it harder to find the best classification for a particular usage scenario.

Previously, an ontology has been proposed for types of user feedback [MPG15], but we do not know of any previous effort that combines classifications for RE in a comprehensive taxonomy in a way that would (1) help to understand the purposes for which user feedback can be classified, and (2) contribute to an initial harmonization of the focus and vocabulary of the research in this area. This is why in this paper we present a preliminary taxonomy of classification categories based on an investigation of existing literature on this topic. We present our taxonomy at this early stage in order to foster discussions among RE and NLP experts, and to get inspiration for further improvements to the taxonomy. This contribution is of an analytic nature as it intends to introduce some degree of order in the proliferation of classifications. It is not meant to impose a standardization governing which classifications to use; on the contrary, we hope to inspire researchers and practitioners to use classifications not previously considered. Through this work, we intend to answer the following research questions:

RQ1: Which classification categories have research publications used to classify requirements-relevant content?

RQ2: How can the identified classification categories be structured into a taxonomy?

RQ3: What are possible analysis purposes for which each category of the taxonomy can be used?

in Section 2, we describe the methodology we applied to answer our research questions, followed by a presentation of the resulting taxonomy in Section 3. Section 4 presents a discussion of possible uses of the taxonomy in practice, and in Section 5, we conclude and provide perspectives on further developing the taxonomy.

2 Method

In this section, we first discuss the approach we employed to identify relevant literature (Section 2.1), followed by a presentation of our methodology for deriving our taxonomy (Section 2.2).

2.1 Systematic Literature Review

Within the scope of a larger benchmarking study, we performed a systematic literature review (SLR) [KC07] to obtain a comprehensive and broad overview of the literature on classifying user feedback. We used an SLR to exclude any potential selection bias and prevent gaps in our research. As part of this effort, we noticed that our set of systematically obtained literature works proposed and used many disjunct classification structures and categories. This finding led us to launching an effort towards harmonizing these classification categories, resulting in the taxonomy presented in this work.

Our SLR protocol specifies the research questions, a search strategy including explicit inclusion and exclusion criteria, and the information to be extracted from the primary research found (cf. [KC07]). We defined the following research questions for the SLR:

- **Overall Objective:** What are the state-of-the-art automated approaches for assisting the task of requirements extraction from user feedback acquired from the crowd, and which NLP techniques and features do they use?
- **Objective 1:** Regarding requirements elicitation from user feedback acquired from the crowd, what are the state-of-art automated approaches for classifying user feedback?
- **Objective 2:** How do such approaches classify user feedback?
  - **Objective 2.1:** What are the different sets of categories in which user feedback is classified?
  - **Objective 2.2:** Which automated techniques are used?
  - **Objective 2.3:** What are the characteristics of the user feedback these approaches aim to classify?

To perform our search, we composed a search string by defining search terms, many of which are common terms from known literature, and tested these in different combinations. We also used previously identified papers to verify whether the search string would correctly find these publications. The final search string was as follows:

((“CrowdRE” OR “Crowd RE”) OR (((“User Review” OR “User Feedback” OR “App Review” OR “Feature Requests” OR “User Opinions” OR “User Requirements”)) AND (Classif* OR Framework OR Tool OR “Text Analysis” OR Mining OR “Feature Extraction”) AND “Requirements Engineering”))

We selected papers according to the eight exclusion criteria (EC) and two inclusion criteria (IC) listed below. A paper meeting one or more ECs was excluded from the selection, while a paper meeting one or more ICs and no ECs was included in the selection.

- **EC1:** The paper is not written in English.
- **EC2:** The paper was published before 2013.[1]
- **EC3:** The work or study is not published in a peer-reviewed venue.
- **EC4:** The paper is not related to RE for software products and/or the title is clearly not related to the research questions.
- **EC5:** The paper does not address the topic of requirements extraction from user feedback analysis.
- **EC6:** The paper proposes a tool or dataset that does not aim to assist a requirements extraction process from online user reviews, or could not be used in this way; for example, recommender systems, information retrieval for search engines, or approaches that link source code changes to bug fixes.
- **EC7:** The paper proposes an approach or tool that does not process textual user feedback. For example, approaches that analyze implicit feedback, process requirements documents, or merely collect user feedback instead of processing it.
- **EC8:** The paper proposes an approach that does not make use of automation because the user feedback analysis is done entirely manually; for example, crowdsourced requirements elicitation.
- **IC1:** The paper proposes an approach for filtering out irrelevant user feedback from raw data, regardless of whether or not this is done using classification techniques.
- **IC2:** The paper proposes an approach for classifying user feedback into default predetermined categories.

We first applied our search string to search for suitable papers in March 2018, using three prominent scientific databases on software engineering research: ACM, Springer, and IEEE Xplore. Exclusion criteria EC1–EC3 were applied directly through database filters. This query returned a combined result of 1,219 papers. After removing duplicates and screening the title and abstract (to which EC4–EC8 and IC1–IC2 were applied), 146 papers remained. These included papers for which the results of our title and abstract analysis were inconclusive, so this number also included papers where we were uncertain whether they matched our selection criteria. Further scanning of the introduction and conclusion sections, to which EC4–EC8 were re-applied, reduced the number of papers for data extraction to a total of 40. This work was performed by the first author of this paper, and the third author cross-checked a random subset to assure the quality of this work. Any disagreements were discussed and resolved. We repeated the query on 18 December 2018 to include papers that had been added over the course of 2018, which resulted in 14 new papers, of which 3 were relevant to our SLR, for a total number of 43 analyzed papers. Due to space restrictions, we present the complete list of primary papers in a separate document[2], and reference them in this document with the notation Pn.

Serving the overall goal of the SLR, i.e., to prepare a benchmarking study, we systematically extracted three major groups of data from the selected papers:

- *Dataset-related information*, such as dataset size in number of entries, object granularity (sentence vs. review), source (e.g., app stores, social media), and mean text object size.
- *NLP techniques applied*, such as algorithms, parsers, ML features, and text pre-processing techniques.
- *Classification categories* into which the tool was designed to classify user feedback, along with their definitions, where available. We also paid specific attention to any explicit rationales behind design decisions made for a tool to understand for which goal or under which circumstances specific categories are best used.

The aggregated overview of the third group of data, “classification categories”, revealed that a benchmarking study would be impeded by the use of different categories. This finding led to our efforts to derive a taxonomy.

2.2 Taxonomy Derivation

We established our taxonomy of user feedback classifications in five steps:

[1] This year was chosen because prior to the introduction of CrowdRE in 2015 [GDA15], the analysis of online user feedback via NLP for RE was not considered to be a serious source of requirements. We additionally considered six years as a technical obsoleteness threshold to fit our paper selection efforts to our resource constraints.

[2] Bibliography of primary studies: zenodo.org/record/2546422, doi:10.5281/zenodo.2546422

Step 1: Collect and complete categories. Having gathered the various classification categories as part of our SLR, we created an overview listing the categories used in each paper, along with their source. We then verified that we had identified all the relevant information from each paper.

Step 2: Merge similar categories. Many of the primary studies presented their own category definitions. To organize them, we inspected their definitions in order to identify similar classification categories that intend to filter the same type of text but have a different name. We then determined the most appropriate name and description for this category. If a paper did not define or explain what the categories used should filter, we assumed that they adopted the same definition as the papers they discussed in their related work section. If any doubt persisted, we contacted the authors by email.

Here is an example of how we merged categories: The category "Feature Request" received this name because it was the most prevalent name in the literature, although it combines the categories "User Requirements" from P20, "Functional Requirements" from P24, "Feature Requests" from P25, and "Request" from P6, all of which we found to refer to texts containing requests for functional enhancements, either by implementing new features or by enhancing existing ones. We then based the definition of this category on the definitions found in P1 and P31. For space reasons, we provide the complete overview of the 78 merged categories[3], along with their definitions and references to the papers in which they were found, in a separate document[4]. The names of all classification categories are also shown in the taxonomy.

Step 3: Group related categories. The studies in P11 and P24 on quality-related aspects used the ISO 25010 software product quality characteristics [ISO10], while P2, P17, P26 and P27 based their work on notable publications from *user experience* (UX) [BAH11, KR08, Nie93] and the ISO 25010 quality-in-use characteristics [ISO10]. These served as the initial framework for clustering our categories because most other papers did not compose the categories or their definitions systematically. Similarity in definitions or even names allowed us to draw parallels to these standardized structures. However, we also took heed not to include characteristics that cannot be found in user feedback according to research, such as "Maintainability" in software product quality, as found in P11. Similarly, we omitted the ISO 25010 quality-in-use characteristics "Freedom from Risk" and "Context Coverage" with their sub-characteristics because these were not found in the UX research, possibly because it might be impossible to estimate them based on the opinion of users. Conversely, we included refinements of these frameworks found in the literature. For example, "Battery" in P4 and P20 refines "Resource Utilization". Relationships between papers, for example papers written by some of the same authors or referencing similar works, were used as indicators that particular categories could be grouped. For example, the 21 categories on UX were found in six papers that aimed to identify UX-related information in user reviews. This is how, in addition to the aforementioned papers on UX, we identified that P26 focuses on a higher-level goal in which the trait of UX is only one classification, while P8 and P25 identify UX traits without further distinguishing them.

After having made attributions based on the standardized framework, we sought to identify patterns among the remaining categories so that they could be organized into conceptually distinct groups. Sentiment-related categories clearly stood out, even though we are aware that some works, such as P11 and P18, juxtapose them with other classification categories. All other categories were initially sorted according to what they aim to filter from the text. Moreover, two works suggested additional categorization structures: types of *topics* was suggested in P25, which we found to be compatible with the ISO 25010 software product quality characteristics, and the author's *intention* was suggested in P35, which we adopted and expanded through discussions with peers. In this way, we came up with the four groupings in our taxonomy and succeeded in to assigning all categorizations to a single group, except for "Learnability", which appears twice in our taxonomy (under Topic and UX) due to its proximity to concepts in both groups.

Step 4: Identify subgroups. Once we had established the four main groups with their categories, we subdivided them into logical subgroups to provide even more structure. For example, we found the category Topic to contain all categories of user feedback that address topics regarding the software product, specifically general statements, particular functions or qualities of the product, or aspects from its extended context.

Step 5: Validate taxonomy. Finally, we performed an early validation of our taxonomy through individual commenting sessions with five domain experts—three RE experts and two UX experts—three of whom have experience in both academia and industry. Their feedback predominantly led to making clearer distinctions or partially reorganizing some clusters of categorizations. The resulting preliminary taxonomy is presented in Section 3.

[3] "Learnability" appears twice in our taxonomy, but is counted once.

[4] Table of classification categories: zenodo.org/record/2577863, doi:10.5281/zenodo.2577863

3 Taxonomy

By grouping the categories identified in the literature, we composed the taxonomy shown in Figure 1. Rounded rectangles in the taxonomy represent classification categories from the literature. One-way arrows signify a subset relationship between categories. Each category except for "Requirements-Irrelevant" is covered under "Requirements-Relevant". Similarly, "Satisfaction" includes "Trust", "Pleasure", "Comfort", and "Utility". Swimlanes within each group represent logical subgroups to further organize the classification categories. Double-sided arrows show explicit antagonists, i.e., categories that cannot be assigned to the same text snippet as a matter of principle.

The premise of this taxonomy builds on the distinction between "informative" content and other, non-informative content that according to P3 does not contribute to RE purposes. We renamed this distinction "Requirements-Relevant" and "Requirements-Irrelevant". The definition we use for "Requirements-Relevant" does not significantly differ from "Informative", but we had to change the scope for the category "Requirements-Irrelevant" because it includes several categories that have been used in the literature to discard certain types of text: "Other", "Noisy", "Unclear", "Unrelated" from P13, "Non-Bug" from P19, and "Miscellaneous and Spam" from P41.

The primary papers proposed a wide range of different classification categories, such as 14 unique categories in P26 and 23 unique categories in P2, while others classified requirements-relevant feedback into just three overall categories, such as "Suggestions for New Features", "Bug Reports", and "Other" in P39. Our taxonomy consists of four groups of user feedback classifications: "Sentiment", "Intention", "User Experience", and "Topic", which we will describe separately in the following subsections. Note that these categorizations are not mutually exclusive, but can also be used in combination, which we will further discuss in Section 4.

3.1 Sentiment

We found several papers on CrowdRE research, P1, P21, P23 and P33, that applied *sentiment analysis*; a commonly applied NLP technique that determines the extent to which texts or elements of such texts are positive or negative. Most sentiment analysis techniques search for predefined sentiment-related words cataloged in dictionaries such as SentiWordNet or AFFIN to assign a word-specific sentiment score on a bipolar scale ranging from very *negative* (e.g., -2) to very *positive* (e.g., +2) to calculate a total score for a sentence or an entire text, like in P33 and P42.

Some techniques merely distinguish between "Positive", "Negative", and "None" (i.e., neutral), like in P1 and P42, treating sentiment analysis as a binary or ternary classification problem. In addition to determining the polarity, some review classification tools used for CrowdRE have suggested classification categories such as "Praise" and "General Complaint", as suggested in P14, which enable them to make a better assessment of how users perceive the product even if only short user feedback is given.

According to P12, associating sentiment analysis with information from other classifications such as Topic (see Section 3.4) can reveal user acceptance levels regarding specific aspects of the software, based on which the aspects receiving the most criticism can be prioritized to be improved first.

3.2 Intention

According to several publications by one research group, P9, P10, P35, and P36, understanding the motivation or drive behind why a user provides feedback can help determine the requirements of this user. This notion underlies the classification according to the user's *intention* or goal, which we could subdivide into requesting, informing, and reporting intention.

Informing user reviews typically seek to persuade or dissuade other crowd members to use the product or to provide a justification for why a particular star rating was given. P13 asserts that users will often describe what was poor or excellent about their interaction with the product. The category "Job Advertisement" may seem unusual in this context, but was used in P14 to classify user feedback on Twitter regarding a job offering at a software company that may be of interest to non-technical stakeholders such as marketing representatives, and for the general public. When informing user feedback also addresses aspects of a product that are present or absent, they may provide interesting topics (see Section 3.4).

Reporting user reviews intend to inform the developer of the product of a problem or defect the user found, which will often be a "Bug Report". Because bug reports are usually objective descriptions of problems and quality issues found in already present features, according to P25 they are a popular user feedback classification

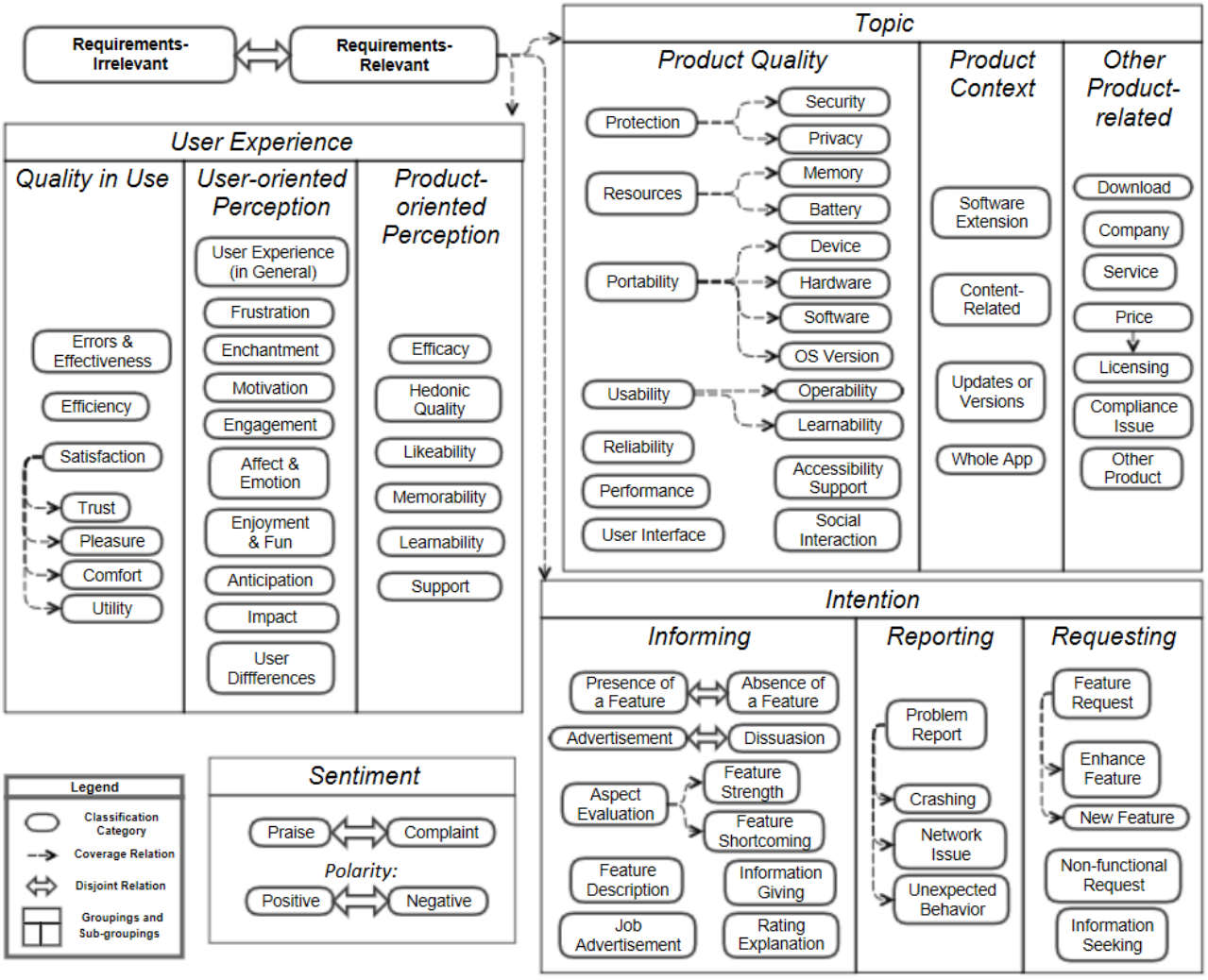

Figure 1: Our preliminary taxonomy for user feedback classification categories.

type for identifying possible functional or quality requirements. For this reason, P26 further subdivides them into categories such as "Network Issue", "Unexpected Behavior", or "Crashing". These will often coincide with classifications of quality aspects (see Section 3.4).

Requesting user reviews harbor the type of user feedback classification found most frequently in the primary studies we analyzed in our work, namely "Feature Requests", which according to P26 represent requests from users to add new (or reintroduce previous) functional aspects to the product, or to remove, modify, or enhance existing features. Users may also make requests to improve a particular quality, for example to make the product faster, more reliable, or more compatible with other systems, or they may place a request to receive information about the product.

3.3 User Experience

An important aspect of user feedback according to P11 is that it is written by users who report on their practical experience with a software product. As a result, aspects of *UX* relate to user requirements because they reflect the users' perceptions of the product or their response to the use or anticipated use of this product. This is why UX and RE are often addressed together in development activities such as elicitation, prototyping, and testing. We found several works, P2, P17, and P27, that sought to classify parts of texts according to UX-related dimensions. According to the ISO 9241-210 standard, the opinions found in user reviews on UX are shaped by a user's personal emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors, and accomplishments that occur before, during, and after use [ISO09]. What distinguishes the classifications in this

group from others is that they are of a more subjective nature [MHBR02]. As a result, a classification regarding a UX aspect does not primarily result in explicit suggestions, but rather provides information about the users' emotions, motivation, and expectations. These may be indicative of problems (e.g., as a source of frustration) or well-liked features (e.g., as a source of excitement).

Several classifications in this group coincide with some of the ISO 25010 *quality-in-use* characteristics [ISO10], specifically "Efficiency", "Effectiveness" (called "Errors and Effectiveness" in P2 and P17), and "Satisfaction" with its four sub-characteristics "Trust", "Pleasure", "Comfort", and "Usefulness". The second and third subgroup in this category sort the classifications into *user-oriented perception*, which involves emotional and behavioral aspects of the user, and *product-oriented perception*, which are opinions that can be attributed to the product or its context. Similar to sentiment analysis (see Section 3.1), the perception of users regarding the UX can provide an indication of product acceptance because greater enjoyment with the product will increase acceptance. Together, according to P2 these analyses can help determine how users react to individual features.

3.4 Topics

The final group of classifications assesses the software product, its aspects, and its context as specific *topics* on which users share their opinion in user feedback. This may reveal actual requirements if the user provided sufficient information. The classifications in this group are distinguished into user opinion pertaining to the functions, quality, or context of a product, and are described in P4, P9, P10, P11, P13, P26, P27, P35, and P36.

Product quality aspects often involve variations of the ISO 25010 software product quality characteristics [ISO10], such as in P4, P11, P26, and P40, with some classifications going into more detail than the standard prescribes (e.g., specifically categorizing user feedback on "Battery" in P4). *Product context* aspects found in P9, P10, P35 and P36 deal with the functional aspects of interoperability, with planned extensions mostly to other software products, discussions of content created or accessible through the product, the behavior of the product in a specific version, and general opinions; the latter do not specify which aspect of the product a user finds good or bad and might does not necessarily pertain to the product itself. *Other product-related* aspects include the users' opinions on the pricing, the development company or team, and the quality of the service they provide, as well as comparisons users make between the product and competitor products to describe what functionality is missing or unique in the product. Classifications on topics may correlate with classifications on user intention (see Section 3.2), especially when users address a product function to make a request, whereas bug reports often address defects in product quality.

4 Practical Application

A key finding of this work is that the types of classifications can be placed into four main groups: Sentiment, Intention, UX, and Topic. The classifications in these four groups of our taxonomy are conceptually different. This not only means that they will produce different results, but also that they may be better suited for some purposes than for others, which we will explore in this section.

From the primary papers and our own experience, we derived seven common RE activities that can benefit from input from user feedback analysis. Most papers pursued only one kind of RE activity, except for some exploratory works, such as P14. The activities are listed in Table 1, where we indicate which of the classification groups are better suited than others. An example of how this overview can be read: The activity of *eliciting requirements* from user feedback is most likely to benefit from a classification according to topic to identify user feedback that addresses the quality, functions, or context of a product. Additionally, assessing user feedback by its intention may reveal *requesting* user feedback and help to specifically find feature requests. Conversely, the Sentiment and UX categories are less suited because they may only lead to requirements indirectly, usually requiring a manual inspection to find them.

Overall, we found that for each activity, two or three groups are suited. Moreover, each of the four classification groups can serve multiple purposes within RE, showing that they are suitable for obtaining different kinds of RE-relevant knowledge, provided users disclose this knowledge in their feedback. These findings can be interpreted in three ways:

- The application of the classification categories of a particular group may be suitable for more purposes than the ones to which they have been applied so far. For example, UX analysis has focused mainly on product acceptance and usage context, but would also be useful for identifying unique selling propositions and potential process improvements.

- One may choose to apply classification according to just one suitable group. This choice may depend on a trade-off between the amount of work required to perform the analysis versus the quality of the results, as some types of analyses are relatively easy to set up (e.g., sentiment analysis), while others can provide deeper insights if more effort is spent on tailoring them to RE.
- This outcome also suggests that the classification groups are not mutually exclusive and can support each other when used in combination. For example, Sentiment and Topic could provide sentiment scores and extracted features, respectively, which could then be aggregated to obtain an overview of the best- and worst-rated product functions.

Table 1: Suitability of classification groups for typical RE activities.

Analysis Goal	Sentiment	Intention	User Experience	Topic
Elicit Requirements		×		×
Measure Product Acceptance	×		×	×
Understand Usage Context		×	×	
Identify Software Problems	×	×		×
Identify and Prioritize Ideas		×		×
Identify Unique Selling Propositions	×	×	×	×
Identify Process Improvements			×	×

5 Conclusion and Outlook

In this paper, we presented a practical auxiliary finding from an SLR into research on the classification of user feedback in CrowdRE research, namely, a preliminary taxonomy of the classification categories found in this research. We found a total of 78 unique classification categories, counting the duplicate occurrence of "Learnability" once (**RQ1**). For space reasons, we had to make the table listing the classification categories available as a separate file[5], but their names are all shown in the taxonomy. Even though the number of unique classification categories was higher than anticipated, it did confirm our suspicion that the lack of an existing structure has caused a proliferation of classification categories in CrowdRE research, which became especially evident from the different names being used for the same concepts. Moreover, several primary papers failed to explain how their categories were chosen or to provide a clear definition of these categories, suggesting that some of the categories found were constructed rather freely. Conversely, only five papers, P2, P11, P17, P24, and P40, based their category definitions entirely on a formal standard. To structure the large number of categories, we took a systematic approach towards establishing a taxonomy (shown in Figure 1), which consists of four main groups: Sentiment, Intention, UX, and Topic (**RQ2**), revealing the four predominant foci of identifying information in user feedback that is of relevance to RE. Finally, we assessed how suitable the classifications of each group are for typical RE-related analyses, which revealed that for most purposes, classifications from different groups can be used (**RQ3**). The choice of classification will often depend on a trade-off between the degree of detail required and the ease of configuring and performing the analysis.

One aspect that was revealed through the taxonomy is that its groups are not mutually exclusive, and that certain aspects can be identified through different classifications; most notably bug reports and feature requests. We argue that this is no contradiction to the way this classification is structured, but rather a logical result of the strong correlation between the categories. For example, although Sentiment is a category of its own, the degree to which user feedback is positive or negative often also underlies the other three groups. We also observe similar overlaps between categories of authoritative standards; for example, according to the ISO 25010 standard [ISO10], poor maintainability during development will likely affect reliability at runtime, with the distinction being the perspective taken. Our taxonomy does not seek to impose a standardization, but rather to be a constructive source of inspiration for research and industry applications. It is also intended as a first step towards introducing harmonization between the kinds of analysis performed and the naming used for the categories.

The premise of this ontology was the bottom-up construction in which we organized the existing classification categories used in the literature. Although it would be possible to theorize about including other potentially useful categorizations in our taxonomy, we present only those categories that research has confirmed to be

[5]Table of classification categories: zenodo.org/record/2577863, doi:10.5281/zenodo.2577863

appropriate for classifying user reviews, omitting those that have been shown or assumed to not be found in user feedback (see Section 2.2 for examples). Moreover, due to the nature of the research, we considered only those categories that have been applied in research studies; an assessment of the categories used in commercial tools available on the market may reveal additional categories. We intend to further validate this taxonomy with specialists in the field of software quality assurance, RE, and UX, and to test its practical applicability as a framework for selecting appropriate classification categories depending on the goal of the user feedback analysis. Furthermore, we believe the taxonomy could be part of a quality framework with guidelines regarding best practices for using classification categories for RE. Such a framework could include metrics for evaluating the quality of classification tools, and the taxonomy could serve as a means for standardizing the classification categories in order to facilitate benchmarking with regard to the quality of the results produced by different tools.

Acknowledgments

We would like to thank the experts, including Dr. Fabiano Dalpiaz, Dr. Jörg Dörr, and Dr. Nash Mahmoud for reviewing earlier versions of the taxonomy. We thank Sonnhild Namingha from Fraunhofer IESE for proofreading this article.

References

[BAH11] Javier A. Bargas-Avila and Kasper Hornbæk. Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2689—2698, 2011.

[Ber17] Daniel Berry. Evaluation of tools for hairy requirements and software engineering tasks. In *Proceedings of the IEEE 25th International Requirements Engineering Conference (RE) Workshops*, pages 284–291, 2017.

[CLH+14] Ning Chen, Jialiu Lin, Steven C. H. Hoi, Xiaokui Xiao, and Boshen Zhang. AR-Miner: Mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering (ICSE)*, pages 767–778, 2014.

[GDA15] Eduard C. Groen, Joerg Doerr, and Sebastian Adam. Towards crowd-based requirements engineering a research preview. In Samuel A. Fricker and Kurt Schneider, editors, *Requirements Engineering: Foundation for Software Quality*, pages 247–253, Cham, 2015. Springer.

[GIG17] Emitza Guzman, Mohamed Ibrahim, and Martin Glinz. A little bird told me: Mining tweets for requirements and software evolution. In *Proceedings of the IEEE 25th International Requirements Engineering Conference (RE)*, pages 11–20, 2017.

[GKH+17] Eduard C. Groen, Sylwia Kocpzyńska, Marc P. Hauer, Tobias D. Krafft, and Joerg Doerr. Users —The hidden software product quality experts? A study on how app users report quality aspects in online reviews. In *Proceedings of the IEEE 25th International Requirements Engineering Conference (RE)*, pages 80–89, 2017.

[GSA+17] Eduard C. Groen, Norbert Seyff, Raian Ali, Fabiano Dalpiaz, Joerg Doerr, Emitza Guzman, et al. The crowd in requirements engineering: The landscape and challenges. *IEEE Software*, 34(2):44–52, March/April 2017.

[GSK+18] Eduard C. Groen, Jacqueline Schowalter, Sylwia Kocpzyńska, Svenja Polst, and Sadaf Alvani. Is there really a need for using NLP to elicit requirements? A benchmarking study to assess scalability of manual analysis. In Klaus Schmid and Paolo Spoletini, editors, *Requirements Engineering – Foundation for Software Quality (REFSQ) Joint Proceedings of the Co-Located Events, CEUR Workshop Proceedings 2075*, 2018.

[IH13] C. Iacob and R. Harrison. Retrieving and analyzing mobile apps feature requests from online reviews. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR)*, pages 41–44, 2013.

[ISO09] ISO. ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. Technical report, ISO, 2009.

[ISO10] ISO/IEC. ISO/IEC 25010 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. Technical report, ISO/IEC, 2010.

[JM17] Nishant Jha and Anas Mahmoud. Mining user requirements from application store reviews using frame semantics. In P. Grünbacher and A. Perini, editors, *Requirements Engineering: Foundation for Software Quality (REFSQ), LNCS 10153*, pages 273–287, Cham, 2017. Springer.

[KC07] B. A. Kitchenham and S Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE–2007–01, School of Computer Science and Mathematics, Keele University, 2007.

[KR08] Pekka Ketola and Virpi Roto. Exploring user experience measurement needs. In *Proceedings of the 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (VUUM)*, pages 23–26, 2008.

[LL17] Mengmeng Lu and Peng Liang. Automatic classification of non-functional requirements from augmented app user reviews. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 344–353, 2017.

[MFH+16] Thorsten Merten, Matúš Falis, Paul Hübner, Thomas Quirchmayr, Simone Bürsner, and Barbara Paech. Software feature request detection in issue tracking systems. In *Proceedings of the IEEE 24th International Requirements Engineering Conference (RE)*, pages 166–175, 2016.

[MHBR02] Andrew Monk, Marc Hassenzahl, Mark Blythe, and Darren Reed. Funology: Designing enjoyment. In Loren G. Terveen and Dennis R. Wixon, editors, *Proceedings of the CHI'02 Extended Abstracts on Human Factors in Computer Systems (CHI EA)*, pages 924–925, 2002.

[MN15] Walid Maalej and Hadeer Nabil. Bug report, feature request, or simply praise? On automatically classifying app reviews. In *Proceedings of the IEEE 23rd International Requirements Engineering Conference (RE)*, pages 116–125, 2015.

[MPG15] Itzel Morales-Ramirez, Anna Perini, and Renata Silva Souza Guizzardi. An ontology of online user feedback in software engineering. *Applied Ontology*, 10(3–4):297–330, 2015.

[Nie93] Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, 1993.

[WM17] Grant Williams and Anas Mahmoud. Mining Twitter feeds for software user requirements. In *Proceedings of the IEEE 25th International Requirements Engineering Conference (RE)*, pages 1–10, 2017.

[WZL+18] Chong Wang, Fan Zhang, Peng Liang, Maya Daneva, and Marten van Sinderen. Can app changelogs improve requirements classification from app reviews?: An exploratory study. In *Proceedings of theACM/IEEE 12th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Article 43, 2018.