A TAXONOMY FOR USER FEEDBACK CLASSIFICATIONS

Rubens Santos, Eduard C. Groen, Karina Villela
What do we want?

A BENCHMARKING of user feedback classification approaches for RE (CrowdRE)

When do we have it?

Now, see...

the differences between the approaches we found actually make it kind of difficult to make a proper comparison that tells us reliably which approaches may be better suited for RE so that we are some several steps away from performing a benchmarking which may require researchers to re-do analyses or to provide us with their data in order for us to perform those analyses ourselves for their results to be comparable on the various levels that these analyses currently differ to such great extents...not now...
The Idea of Our Benchmarking is Simple…

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✅</td>
</tr>
<tr>
<td>2</td>
<td>✅</td>
</tr>
<tr>
<td>3</td>
<td>✅</td>
</tr>
<tr>
<td>4</td>
<td>✅</td>
</tr>
<tr>
<td>5</td>
<td>✅</td>
</tr>
</tbody>
</table>

Wrench image source: [MidnightLightning / WikiMedia Commons](https://commons.wikimedia.org/wiki/File:Wrench.png), public domain
…The Reality of this Benchmarking is Difficult…

**Algorithms** are used in combination with different combinations of other **NLP techniques**, including primary and secondary machine learning features, semi-supervised classification algorithms, and pre-processing techniques.

**Datasets** differ, among other things, in size (number of entries), object granularity (sentence vs. review), sources covered (e.g., app stores, social media), and mean text object size.

**Analyses** use different classification categories according to different definitions and gold standards.
Datasets differ, among other things, in size (number of entries), object granularity (sentence vs. review), sources covered (e.g., app stores, social media), and mean text object size.

Algorithms are used in combination with different combinations of other NLP techniques, including primary and secondary machine learning features, semi-supervised classification algorithms, and pre-processing techniques.

Hurdle 2: An overview of user feedback classification approaches

Analyses use different classification categories according to different definitions and gold standards

Hurdle 1: A taxonomy for user feedback classifications

Focus of this presentation

Benchmarking

Systematic Literature Review

Hurdle 3 and further:
• Comparing datasets
• Assessing the influence of NLP techniques
• Aligning analyses
• Etc.

...But We Are Doing This Benchmarking
Systematic Literature Review

- Conducted according to Kitchenham, with an SLR protocol specifying:
  - objectives / research questions,
  - a search strategy with inclusion/exclusion criteria & a search string,
  - a data extraction strategy.

✔ Note: The SLR is not the main focus of this presentation!
- We’re showing a “byproduct” in a preliminary form
  - Focusing only on the first hurdle that we had to overcome
  - We wanted to get this material out there, so you can work with it!
SLR: Objectives

**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?
SLR: Paper Search

Performed March 2018 (+ December 2018)

1,219 papers
  + 14 papers
• Found by search string

146 papers
• After filtering by title & abstract

40 papers
  + 3 papers
• After filtering by intro & conclusion

**EC1**: not English
**EC2**: before 2013
**EC3**: not peer-reviewed

**IC1**: filters out irrelevant user feedback
**IC2**: classifies into predetermined categories

**EC4**: not RE / unrelated title
**EC5**: not on req. extraction from user feedback
**EC6**: tool not (usable) for requirement extraction
**EC7**: tool does not process textual user feedback
**EC8**: manual processing without automation
SLR: Data Extraction from 43 Papers

1. Dataset-related information
   - E.g., dataset size in number of entries, object granularity, sources, mean text object size

2. NLP techniques applied
   - E.g., algorithms, parsers, ML features, text pre-processing techniques

3. User feedback classification categories → Taxonomy
   - E.g., name, definition, rationale/goal
Hurdle for Benchmarking

- Papers propose/use many disjunct classification structures and categories
  - Need for harmonization
    - Taxonomy
Taxonomy Composition
Step 1: Collect and Complete Categories

- Overview of classification categories (name, definition, source)
- Verification step that all relevant information was collected

**Note:** Our approach is descriptive; we include all categories that:
- Are used in the literature
- Have garnered useful results in user feedback

**Examples:**
- The ISO 25010 software product quality characteristic “Maintainability” was not found in user feedback → exclude
- “Freedom from Risk” & “Context Coverage” were omitted from papers on the ISO 25010 quality-in-use characteristics → exclude
- “Job Advertisement” was used in literature → include
Taxonomy Composition
Step 2: Merge Similar Categories

- Harmonization of categories by definition
  - Merging categories that intend to filter the same type of text, even if they have a different name
  - Determining the most appropriate name and description for this category

- Example: “Feature Request”
  - Requests for functional enhancements
  - Most prevalent name in the literature is “Feature Request”
    - In some papers “User Requirements”, “Functional Requirements” or “Request”
  - Definition was based on papers P1 and P31
Taxonomy Composition
Step 3: Group Related Categories 1/5

- We sorted papers by their compatibility of categories
- Mutually exclusive labels, multi-label, functional labels only
- Sentiment did not really fit anywhere → omit?

<table>
<thead>
<tr>
<th>Index</th>
<th>Category</th>
<th>Mutually Exclusive Categories?</th>
<th>Mutually Exclusive Categories (not multi-label)</th>
<th>Input Granularity</th>
<th>Mutually Exclusive Categories?</th>
<th>Mutually Exclusive Categories (not multi-label)</th>
<th>Input Granularity</th>
<th>Sentiment (should be excluded)</th>
<th>Input Granularity</th>
<th>Sentiment (should be excluded)</th>
<th>Input Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>Functional dysfunctional</td>
<td>Yes</td>
<td>30 functional bug, functional demand, non-functional request, other</td>
<td>4 review</td>
<td>functional bug, functional demand, non-functional request, other</td>
<td>4 review</td>
<td>positive, negative</td>
<td>3 review</td>
<td>functional bug, functional demand, non-functional request, other</td>
<td>4 review</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>Requirement or non-requirement</td>
<td>Yes</td>
<td>37 Functional Requirements (compatibility, usability, portability, performance), Functional Requirements (FRs), Others</td>
<td>2 review</td>
<td>37 Functional Requirements (compatibility, usability, portability, performance), Functional Requirements (FRs), Others</td>
<td>2 review</td>
<td>positive/negative</td>
<td>2 review</td>
<td>37 Functional Requirements (compatibility, usability, portability, performance), Functional Requirements (FRs), Others</td>
<td>2 review</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Information, non-information</td>
<td>No</td>
<td>61 Sentiment (should be excluded)</td>
<td>1 review</td>
<td>61 Sentiment (should be excluded)</td>
<td>1 review</td>
<td>Yes</td>
<td>1 review</td>
<td>61 Sentiment (should be excluded)</td>
<td>1 review</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Improvement request and other</td>
<td>No</td>
<td>92 Improvement Request and Other</td>
<td>2 review</td>
<td>92 Improvement Request and Other</td>
<td>2 review</td>
<td>positive/negative</td>
<td>2 review</td>
<td>92 Improvement Request and Other</td>
<td>2 review</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>Feature request, other</td>
<td>Yes</td>
<td>115 Feature request, other</td>
<td>4 review</td>
<td>115 Feature request, other</td>
<td>4 review</td>
<td>positive/negative</td>
<td>4 review</td>
<td>115 Feature request, other</td>
<td>4 review</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>Bug, nondrug</td>
<td>Yes</td>
<td>120 Bug report, nondrug</td>
<td>2 review</td>
<td>120 Bug report, nondrug</td>
<td>2 review</td>
<td>positive/negative</td>
<td>2 review</td>
<td>120 Bug report, nondrug</td>
<td>2 review</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>Bug report, other</td>
<td>Yes</td>
<td>104 Bug report, other</td>
<td>3 review</td>
<td>104 Bug report, other</td>
<td>3 review</td>
<td>positive/negative</td>
<td>3 review</td>
<td>104 Bug report, other</td>
<td>3 review</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>User requirements, bug report or other</td>
<td>No</td>
<td>6 User requirements, bug report or other</td>
<td>1 review</td>
<td>6 User requirements, bug report or other</td>
<td>1 review</td>
<td>positive/negative</td>
<td>1 review</td>
<td>6 User requirements, bug report or other</td>
<td>1 review</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Information, non-information</td>
<td>No</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td>positive/negative</td>
<td>1 review</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Information, non-information</td>
<td>No</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td>positive/negative</td>
<td>1 review</td>
<td>4 Information, non-information</td>
<td>1 review</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Information, non-information</td>
<td>No</td>
<td>6 Information, non-information</td>
<td>1 review</td>
<td>6 Information, non-information</td>
<td>1 review</td>
<td>positive/negative</td>
<td>1 review</td>
<td>6 Information, non-information</td>
<td>1 review</td>
<td></td>
</tr>
</tbody>
</table>
Taxonomy Composition
Step 3: Group Related Categories 2/5

**Realization:** Classification is primarily concerned with:

- **"Is this text snippet relevant from an RE perspective?"**
  - If **YES:** Classify as relevant in some way
    - Either into one category, or several categories
  - If **NO:** Discard

![Diagram](attachment://taxonomy_diagram.png)
Taxonomy Composition
Step 3: Group Related Categories 3/5

Initial basis for framework

- ISO 25010 software product quality (P11, P24)
- Existing categories in **user experience (UX)** research (P2, P17, P26, P27)
- ISO 25010 quality-in-use (P2, P17, P26, P27)
Taxonomy Composition
Step 3: Group Related Categories 4/5

Indications for grouping categories 1/2

- Refinements of framework components
  - E.g., “Battery” refines ISO 25010 “Resource Utilization”

- Relationships between papers
  - E.g., same authors, references to similar work (especially UX)

- Patterns
  - E.g., what do the categories aim to filter from the texts?
Indications for grouping categories 2/2

- Suggestions for grouping
  - Maalej and Nabil (2015) suggested types of topics
  - Compatible with ISO 25010 software product quality
  - Panichella et al. (2016) suggested author’s intention
### Taxonomy Composition

**Step 4: Identify Logical Subgroups for More Structure**

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Intention</th>
<th>User Experience</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Quality in Use</td>
<td>• Product Quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• User-oriented Perception</td>
<td>• Product Context</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Product-oriented Perception</td>
<td>• Other Product-related Aspects</td>
</tr>
<tr>
<td></td>
<td>• Requesting</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Informing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Reporting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Taxonomy Composition
Step 5: Perform an Early Validation

- Individual commenting sessions
  - Five domain experts
    - 3 RE, 2 UX; 3 experienced in academia + industry
  - Result: clearer distinctions or partial cluster reorganizations
Taxonomy for User Feedback Classification Categories

- **User Experience**
  - Quality in Use
  - User-oriented Perception
  - Product-oriented Perception
  - User Differences

- **Product Quality**
  - Protection
  - Privacy
  - Security
  - Resources
  - Memory
  - Battery
  - Device
  - Portability
  - Hardware
  - Software
  - OS Version
  - Usability
  - Operability
  - Learnability
  - Usability
  - Reliability
  - Performance
  - Accessibility
  - Support

- **Product Context**
  - Software Extension
  - Content-Related
  - Price
  - Updates or Versions
  - Whole App
  - Other Product-related

- **Intention**
  - Informing
  - Absence of a Feature
  - Presence of a Feature
  - Advertisement
  - Dissuasion
  - Aspect Evaluation
  - Feature Strength
  - New Feature
  - Feature Shortcoming
  - Information Giving
  - Job Advertisement
  - Rating Explanation

- **Reporting**
  - Problem Report
  - Crashing
  - Network Issue
  - Unexpected Behavior
  - Non-functional Request
  - Information Seeking

- **Sentiment**
  - Praise
  - Complaint
  - Positive
  - Negative

- **Legend**
  - Classification Category
  - Coverage Relation
  - Disjoint Relation
  - Groupings and Sub-groupings

© Fraunhofer IESE
Assumption: Sentiment helps determine how users feel about the product
- Usually in the form of classic sentiment analysis
  - Polarity (positive, negative), sometimes intensity
  - Categorization into “Praise” and “General Complaint” (P14)
  - Assesses user perception even with short user feedback

- Sentiment is especially useful to be used in combination with other groups from the taxonomy
Assumption: Understanding why a user provides feedback helps determine the requirements (P9, P10, P35, P36)

Informing: Persuade / dissuade other crowd members, or to justify why a particular star rating was given

Reporting: Point out a problem or defect to the developer

Requesting: Requests to add new / reintroduce previous functional aspects; remove, modify, or enhance existing features or qualities

“Job Advertisement” classifies 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 (P14)
Assumption: Users provide user feedback based on their practical (user) experience with the product

- Therefore: aspects of UX relate to RE

Opinion based on the user’s perception and their response to the (anticipated) use of the product → inherently ambiguous

- Emotions, motivation, expectations

Especially helpful to determine degree of product / feature acceptance
Assumption: Users share their opinion on specific (requirement-related) topics
May reveal requirements if the user provided sufficient information

Product quality aspects for quality requirements (cf. ISO 25010)

Product context on interfaces, accessible content, behavior in a particular version, or general opinions

Other product-related aspects: Users' opinions on the pricing, company (including developers or service), comparison to competitor products
Applying the Taxonomy

- We share our taxonomy so it **may help you**
- Inspire you to consider existing groups for other purposes than those for which you used them so far
- Inform a decision on a time-intensive analysis with high-quality results vs. a quicker but less thorough outcome
- Suggest the use of multiple groups in combination

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

<table>
<thead>
<tr>
<th>Analysis Goal</th>
<th>Sentiment</th>
<th>Intention</th>
<th>User Experience</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elicit Requirements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure Product Acceptance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand Usage Context</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify Software Problems</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify and Prioritize Ideas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify Unique Selling Propositions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify Process Improvements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Keep in Mind:

- The taxonomy is **preliminary**
  - Seeks to be a source of *inspiration* for research and industry applications; not to impose a standardization
  - Suggests a possible *harmonization* between the kinds of analysis performed and the naming used for the categories
  - More validation & testing of its practical applicability needed

- The taxonomy is **descriptive** at this point
  - We organized the *existing* classification categories from the literature
  - No analysis yet of potential categories that theoretically could be useful, or that are used in commercial tools on the market
  - Categories with different names were merged
  - “Learnability” appears twice (counted as once)
Key Findings

- Our preliminary taxonomy of user feedback classification categories for RE (CrowdRE) consists of **four groups with 78 categories**
- Lack of a structure caused a proliferation of categories
  - Contributed by providing a **harmonization**
- Many RE-related purposes for user feedback, thus many categories
  - Focus on what the user finds important in their **intention**, **experience** and **topics** addressed; supported by **sentiment**
- The various groups differ in degree of detail, ease of configuring and conducting the analysis
  - For most purposes, classifications from different groups can be used
    - Strong correlation makes them complimentary
    - Similar to, for example, ISO 25010 software product quality
Future Work

- Validating and further ripening of the taxonomy
  - Possible prescriptive expansion (challenge: needs validation too)
  - Assessment of existing commercial tools’ classifications

- Contribution to our benchmarking
  - Provides a structured, harmonized framework facilitating comparisons
  - Could suggest metrics for comparing and evaluating the quality of classification tools according to the same structure
  - Could support guidelines in a larger quality framework for classification in RE (CrowdRE)
Thank you! (For now...)

[Diagram with various categories and subcategories related to user experience, product quality, and intention.]

Legend:
- Classification Category
- Coverage Relation
- Disjoint Relation
- Groupings and Sub-groupings

Sentiment:
- Praise ↔ Complaint
- Polarity: Positive ↔ Negative

Topic:
- Product Quality
  - Security
  - Privacy
  - Memory
  - Battery
  - Device
  - Software
  - OS Version
  - Operability
  - Learnability
- Product Context
  - Software Extension
  - Content-Related
  - Updates or Versions
  - Whole App
  - Licensing
  - Compliance Issue

Other Product-related:
- Download
- Company
- Service
- Price
- Other Issue