A Quantitative Characterization of Design Thinking Artifacts in Requirements Engineering Education

Takumi Katsuie[1], Shinpei Ogata[1], Kozo Okano[1], Yukako Iimura[2] and Shinobu Saito[2]

[1]Faculty of Engineering Shinshu University, Nagano, Japan
[2]Computer&Data Science Laboratories NTT, Inc., Tokyo, Japan

Abstract

This paper presents a quantitative analysis of emotional expressions in design thinking artifacts and examines their relationship with the strength of emotional expression in requirements articulation and requirement–function continuity. We collected artifacts produced by 76 students in a university software engineering course, including Personas, Service Scenarios, Customer Journey Maps (CJMs), User Story Mappings (USMs), and User Interface Designs, and focused on textual artifacts. Text blocks were manually annotated for requirements and functions, and sentiment intensity was computed using a BERT-based, polarity-independent RMS score. The results showed that sentiment intensity varies across artifacts, correlates with requirements in Personas but not in CJMs, and is higher for requirements associated with multiple functions. These findings provide a foundation for data-driven support in requirements engineering education. The sentiment intensity may serve as a lightweight cue for identifying how students articulate user concerns and transition from user understanding to functional design.

Keywords

Software engineering, Requirements Engineering, Requirements management, Sentiment analysis,

1. Introduction

In modern software development, success depends not only on technical correctness but also on addressing users' needs, contexts, and experiences. Design thinking has therefore gained importance as an approach that emphasizes user understanding and problem framing [1]. Developers externalize this understanding through design thinking methods, which are typically materialized in artifacts capturing user goals, behaviors, emotions, and candidate functionality. Prior work qualitatively discussed the roles of such artifacts in fostering empathy and shared understanding [2, 3, 4, 5, 6, 7]. However, empirical evidence on how emotional characteristics are expressed in these artifacts and how they relate to requirements articulation and cross-artifact continuity remains limited.

In parallel, sentiment analysis has been widely applied in software engineering research to analyze opinions and emotions in textual data [8, 9, 10, 11, 12], yet their application to requirements-related design thinking artifacts has received little attention.

Design thinking spans exploratory activities focused on users' goals and emotions and structuring activities that organize these insights into tasks and functions. As a result, artifacts differ in structure and in how emotional expression appears. However, despite the importance of emotions in early user understanding, their quantitative characteristics across artifacts and their relation to requirements remain underexplored. This study provides an initial quantitative analysis of emotional expressions in design thinking artifacts within the context of requirements engineering education. From an educational perspective, quantitative cues that help instructors efficiently review diverse student-created artifacts remain limited. Sentiment intensity may serve as a scalable proxy for gauging how students articulate user concerns and transition from user understanding to functional design.

Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2026. Poznan, Poland, March 23-26, 2026

24w6020j@shinshu-u.ac.jp (T. Katsuie); ogata@cs.shinshu-u.ac.jp (S. Ogata); okano@cs.shinshu-u.ac.jp (K. Okano); yukako.iimura@ntt.com (Y. Iimura); shinobu.saito@ntt.com (S. Saito)
0000-0001-6996-3073 (S. Ogata); 0009-0006-9865-8362 (K. Okano); 0009-0006-3030-3442 (Y. Iimura); 0000-0002-6259-3521 (S. Saito)

This study analyzes emotional expressions embedded in design thinking artifacts created by 76 students in a university software engineering course, spanning Personas, Service Scenarios, Customer Journey Maps (CJMs), User Story Mappings (USMs), and User Interface (UI) designs. Text blocks were manually annotated for requirements and functions, and the sentiment intensity score was computed using a BERT-based sentiment analysis model.

We address the following research questions:

- **RQ1: How does sentiment intensity vary across artifacts and CJM elements?**
- **RQ2: Does sentiment intensity differ between requirement-related and non-related text?**
- **RQ3: How does sentiment intensity relate to requirement–function mappings?**

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes the course context and dataset, Section 4 presents the methodology, Section 5 reports the results, Sections 6 discuss implications and threats to validity, and Section 7 concludes the paper.

2. Design thinking and Sentiment analysis

In software and service design, design thinking artifacts such as Persona, Service Scenario, CJM, and USM are widely used to support user-centered requirements exploration in early development phases. Practitioner-oriented literature characterizes these artifacts as qualitative and communicative tools for fostering empathy and shared understanding through narrative descriptions of user goals, behaviors, and emotions, as well as holistic representations of user experiences across touchpoints and processes [2, 3, 4, 5, 6, 7]. However, despite their established conceptual roles, there is limited empirical evidence on how these roles are reflected in the textual characteristics of the artifacts themselves, particularly with respect to emotional expression across artifact types and internal components.

In parallel, sentiment analysis has been increasingly applied in software engineering research to analyze affective signals in textual artifacts such as user reviews and development communications [8, 9, 10, 11, 12]. Requirements engineering studies have also explored natural language processing techniques for extracting requirements and establishing traceability links across artifacts [13, 14]. However, these approaches rarely focus on design-thinking-specific artifacts produced during upstream ideation, nor do they examine emotional characteristics in relation to requirements articulation and cross-artifact continuity.

3. Course Context and Dataset

3.1. Course Design

We analyzed artifact sets produced by students during exercises on understanding design thinking methodologies in a user-centered requirements analysis course for third-year undergraduates. After teaching various design thinking methodologies, students practiced applying them to a self-chosen service idea individually without third-party reviews. In the exercise, students sequentially analyzed and created five artifacts: a Persona, a Scenario, a CJM, a USM, and a UI design. These methods capture complementary perspectives of user-centered design, ranging from user characterization and interaction narratives to experience mapping, task structuring, and interface design.

76 students, primarily third-year undergraduates in science and engineering with prior programming and modeling training(e.g., UML), attended the course. We only use artifact sets for which students have explicitly granted permission.

3.2. Analysis Dataset

We extracted textual content from artifacts at the level of semantically meaningful text blocks—covering persona attributes, scenario steps, CJM items (user behaviors, touchpoints, emotions, insights), and

individual user stories in USMs and focused our analysis on four textual artifacts (excluding UI designs, which are predominantly visual). In total, we extracted 3,726 text blocks from 76 artifact sets: 453 from personas, 512 from service scenarios, 1,530 from CJMs, and 1,231 from USMs.

We identified requirements and functions from text blocks, assigned consistent IDs to identical requirements and functions across artifacts, and linked functions to the requirements they satisfy. In the initial trial, a practitioner (hereafter referred to as an annotator) proficient in analyzing software development artifacts and manuals, annotated one deliverable set. After the authors and annotator jointly reviewed the results to establish extraction rules (what to extract and what not to, along with the rationale) and extraction considerations. Using these refined rules, the annotator annotated all deliverables with Microsoft Copilot (ChatGPT-5) and conducted final checks.

4. Methodology

4.1. Sentiment Analysis Model

We fine-tuned a pre-trained regression model using the Hugging Face Transformers library. The model predicts a continuous sentiment polarity score from -1.0 (strong negative) to +1.0 (strong positive), with 0 indicating neutral sentiment.

For fine-tuning, we used Version 2 of the WRIME dataset [15], a large-scale Japanese sentiment analysis dataset constructed from social media posts. WRIME provides both subjective annotations (reflecting the writer's expressed emotions) and objective annotations (reflecting how readers perceive the writer's emotions). In Version 2, sentiment polarity is annotated on a five-point scale from -2 (strong negative) to +2 (strong positive). The dataset contains 35,000 posts, split into 30,000 training, 2,500 validation, and 2,500 test samples. We used the full training split and adopted avg_readers.sentiment, the average polarity score assigned by three independent annotators, as the regression target. These labels were normalized to the range [-1.0, 1.0] by dividing by two.

As the base model, we employed tohoku-nlp/bert-base-japanese-v3, which was pre-trained on the Japanese portion of the CC-100 corpus and Japanese Wikipedia [16]. The model was instantiated using AutoModelForSequenceClassification with num_labels=1, adding a single linear layer for scalar regression. Training was performed using Mean Squared Error (MSE) loss. To explicitly constrain the output range, we applied a tanh activation function to the model's raw prediction during both training and inference. The model was fine-tuned for three epochs with standard hyperparameters for BERT-based regression, including a learning rate of 2×10^{-5} and Mean Squared Error (MSE) loss. The maximum sequence length was set to 512 tokens.

4.2. Sentiment intensity score

Using the fine-tuned model, we assigned sentiment scores to all extracted text blocks. We adopt sentiment intensity as a polarity-independent measure to capture the strength of emotional expression, reflecting how emotions are used to articulate user concerns in early requirements work. Because text blocks often contain multiple sentences with mixed polarity, sentiment intensity provides a robust representation of emotional magnitude that is not affected by cancellation effects. Specifically, each text block was split into sentences, the sentiment model was applied to each sentence to obtain a polarity score $s_i \in [-1, 1]$, and the scores were aggregated using the Root Mean Square (RMS):

Formally, for a text block consisting of n sentences, the sentiment intensity score is defined as:

$$\text{sentiment\_score} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i^2}.$$

This aggregation prevents cancellation between opposing sentiments and captures the overall emotional strength of each text block.

Table 1
Summary of requirement–function relationships

Category	Count	Description
Requirements (total)	1,325	Distinct requirement IDs
None ($N_f = 0$)	410	No linked functions
Single ($N_f = 1$)	499	Linked to one function
Multi ($N_f \geq 2$)	416	Linked to multiple functions
Functions (total)	1,814	Distinct function IDs
Single-req	517	Linked to one requirement
Multi-req	393	Linked to multiple requirements
No-req	904	No linked requirements

4.3. Requirement–function relationships

In this study, cross-artifact continuity is operationalized through requirement–function traceability. To analyze the relationship between sentiment intensity and requirement–function traceability, we examined how sentiment scores vary with the number of functions per requirement. Because the exact requirement may appear in multiple text blocks across different artifacts, we aggregated data at the Requirement ID level. For each requirement ID, we computed a representative sentiment score as the mean of the sentiment intensities across all associated text blocks. We then identified the unique set of linked functions to determine the function count, avoiding duplicates. Based on this count (N_f), requirements were classified into three categories: None ($N_f = 0$), indicating no linked functions; Single ($N_f = 1$), indicating exactly one linked function; and Multi ($N_f \geq 2$), indicating associations with multiple functions. Across all artifacts, we identified 1,325 distinct requirements and 1,814 distinct functions, exhibiting substantial variation in requirement–function relationships (Table 1). We compared sentiment scores across these groups using pairwise Welch's t-tests. To account for multiple comparisons, we applied the Bonferroni correction and adjusted the significance threshold accordingly.

5. Results

This section reports the results of the analyses conducted to address RQ1–RQ3.

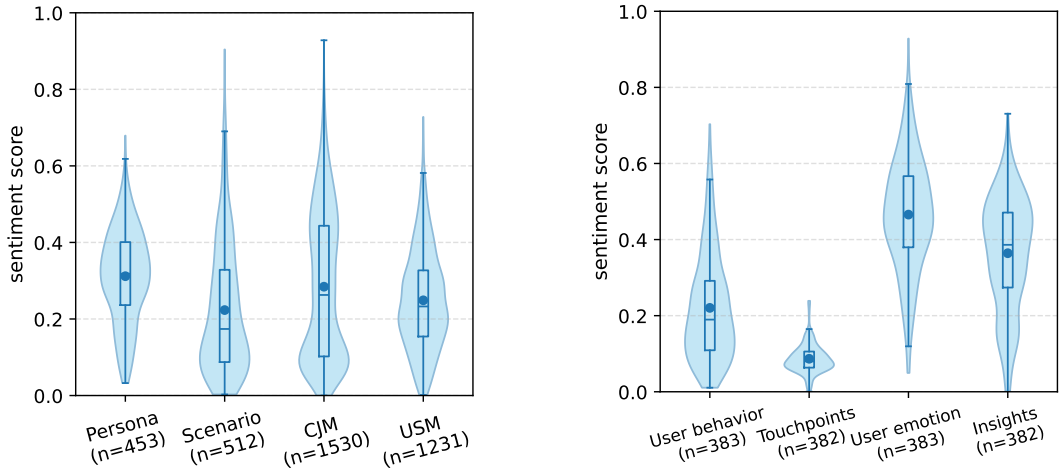
5.1. RQ1: How does sentiment intensity vary across artifacts?

We analyzed sentiment intensity distributions across artifact types (Fig. 1 (a)) and CJM elements (Fig. 1 (b)). Sentiment intensity was observed in all artifact types, but distribution patterns differed markedly (Fig. 1 (a)). Personas exhibit the highest mean sentiment intensity, with density concentrated in the moderate range (0.3–0.4), indicating frequent use of emotionally expressive language. In contrast, Service Scenarios and CJMs are dominated by near-neutral sentiment (0.0–0.1), although CJMs display a wider range with a noticeable tail toward higher sentiment values. USMs show a relatively narrow distribution centered around low-to-moderate sentiment intensity.

Item-level analysis of CJMs (Fig. 1 (b)) reveals clear differences among components. User behavior and Touchpoints were concentrated in low sentiment ranges, while User emotion and Insights exhibited substantially higher sentiment intensity, with User emotion showing the widest distribution. This combination of low- and high-intensity items explains the wide aggregate sentiment range observed in CJMs.

5.2. RQ2: Does sentiment intensity differ between requirement-related and non-related text?

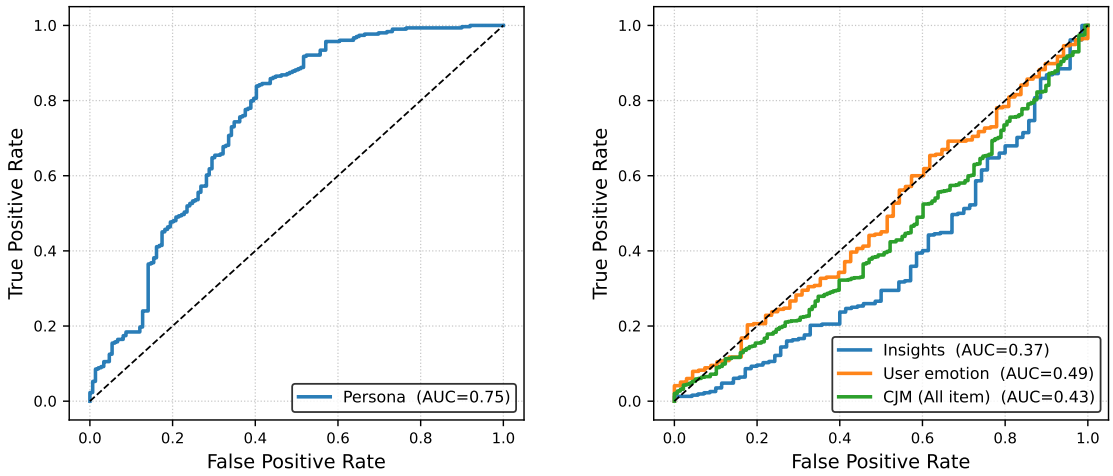
We compared sentiment intensity between text blocks with and without requirement descriptions in Personas and CJMs. Fig. 2 (a) and Fig. 2 (b) report ROC-based discriminative performance.



(a) by artifact type

(b) Item-level within Customer Journey Maps

Figure 1: Distribution of sentiment intensity across design thinking artifacts and CJM items.



(a) Persona

(b) Customer Journey Map

Figure 2: ROC curve for predicting requirement presence from sentiment intensity in Personas.(AUC = 0.75).

In Personas, requirement-related blocks tend to concentrate in the moderate range (0.3–0.4), whereas non-related blocks are more frequent in lower ranges (0.1–0.2); Consistent with this distributional difference, sentiment intensity showed strong discriminative performance for requirement presence in Personas (AUC = 0.75) (Fig. 2 (a)). in CJMs, the distributions largely overlap, and ROC-based discrimination is reported in Fig. 4. In contrast, CJM text blocks with and without requirements exhibited similar sentiment distributions, and sentiment intensity showed limited discriminative power (overall AUC = 0.43; item-level AUCs = 0.37–0.49) (Fig. 2 (b)).

These results indicate that sentiment intensity is informative for identifying requirements in Personas, but not in CJMs. We additionally examined the relationship between sentiment intensity and the presence of function descriptions in Service Scenarios, the User behavior item in CJMs, and USMs. Across all three artifact types, text blocks with and without functions showed substantial overlap in sentiment distributions, with no concentration of function descriptions in higher sentiment ranges. This indicates that sentiment intensity does not serve as a discriminative signal for functions.

5.3. RQ3: How does sentiment intensity relate to requirement–function mappings?

We analyzed the relationship between sentiment intensity and variations in requirement–function mapping (Fig. 3). Statistical testing revealed a significant difference only between the Multi group ($N_f \geq 2$) and the None group ($N_f = 0$), with the Multi group exhibiting higher mean sentiment intensity ($p < 0.0167$, Bonferroni-adjusted). No significant differences were observed between the None and Single groups or between the Single and Multi groups.

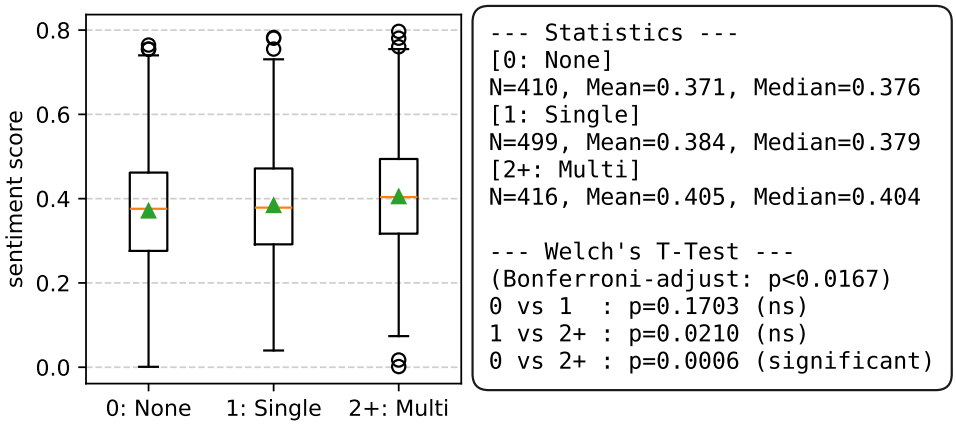


Figure 3: Sentiment intensity by requirement–function mapping type.

6. Discussion

6.1. Roles of emotional expression in design thinking artifacts (RQ1)

Our results show that emotional expressions, measured by sentiment intensity, are unevenly distributed across design thinking artifacts and reflect their intended roles. Personas exhibit higher sentiment intensity, whereas Scenarios and USMs remain largely near-neutral, suggesting an emphasis on objective descriptions of interactions, structure, and scope. CJMs show a mixed pattern: sentiment intensity is higher in *User emotion* and *Insights*, while *User behavior* and *Touchpoints* remain primarily descriptive.

These differences arose from the dual nature of design thinking and the positioning of each artifact within the requirements analysis process. In exploratory activities, emotional expression is important for understanding user goals and motivations, leading to high sentiment intensity in personas and early scenarios. In contrast, in design activities, emotional expression is suppressed in behavioral and touchpoint items of USMs and CJMs to organize user issues into functions and tasks. Thus, the distribution of sentiment intensity corresponds to the roles of artifacts and their phases in requirements analysis.

6.2. Distinct emotional characteristics of requirements and functions (RQ2)

The relationship between sentiment intensity and requirements varies by artifact type. In Personas, sentiment intensity correlates with requirement presence and can discriminate requirement-related text from non-related text, suggesting that emotional expression supports articulating and motivating user concerns during requirements articulation. In contrast, CJMs show no such relationship: sentiment remains high in *User emotion* and *Insights* regardless of whether a requirement is explicitly stated. Across artifact types, sentiment intensity shows no meaningful association with function descriptions, which are consistently expressed in neutral, objective language.

This difference reflects the differing roles of artifacts. Personas emotionally express user frustrations and desires, promoting requirement formation. CJMs separate emotional experiences from requirement

descriptions, making sentiment intensity an unreliable indicator of requirements.

6.3. Sentiment intensity and requirement–function mappings (RQ3)

Sentiment intensity is also associated with requirement–function mapping variation. Requirements linked to multiple functions exhibit higher sentiment intensity than requirements without functional realization, suggesting that emotionally expressive requirements may correspond to broader or more complex user concerns that require decomposition into multiple functional elements.

Therefore, sentiment intensity can be used as a new educational indicator to assess requirement maturity and the need for decomposition. While sentiment intensity may serve as a weak signal indicating design "hot spots," it is insufficient as a sole indicator for single-function requirements.

6.4. Implications for education and practice

In educational settings, sentiment intensity can be used to evaluate the depth of user problem expression and requirement maturity, enabling efficient feedback. The strong correlation between sentiment intensity and requirements in personas indicates whether students deeply engage with user motivations, enabling instructors to identify personas that need further elaboration or clarification. Since sentiment intensity depends on artifact roles, applying it uniformly across heterogeneous artifacts can lead to misunderstandings. In CJMs, it is important to teach the distinction between emotional experiences and requirement descriptions to prevent misinterpretation of emotional reactions as requirements. Insights on requirement–function mappings can trigger instructions on decomposing broad issues into functions.

These implications are particularly useful in large classes handling numerous diverse artifacts, where sentiment intensity can reduce review workload while supporting focused feedback as a scalable metric.

6.5. Validity considerations

We acknowledge several limitations. The dataset is limited to student-generated artifacts from a single course, which may limit external validity. Using a single Japanese BERT-based sentiment model may introduce bias; future work should explore multiple models and languages to assess generalizability. Finally, expert annotation without quantified inter-annotator agreement may introduce subjectivity, and future studies should incorporate multiple annotators and report agreement metrics to strengthen reliability.

7. Conclusion

This study conducted a quantitative investigation of emotional expressions in design thinking artifacts and explored how these expressions relate to requirements articulation and requirement–function continuity. The results showed that sentiment intensity varies across artifact types and internal components, correlates with requirements in personas but not in customer journey maps, and shows no association with function descriptions. Requirements linked to multiple functions also exhibited higher sentiment intensity than those without functional realization.

This study provided an initial quantitative analysis of emotional expression in design thinking artifacts and their relationship to requirements. The findings offer a foundation for data-driven support in requirements engineering education, suggesting that sentiment intensity can serve as a lightweight cue for how students articulate user concerns and transition from user understanding to functional design. Future work should examine longitudinal transitions across artifacts and explore automated feedback mechanisms for RE education.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check, Paraphrase, and reword. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Brown, B. Katz, Change by design, Journal of Product Innovation Management 28 (2011) 381–383.

[2] T. Brown, Design thinking, Harvard Business Review 86 (2008) 84–92.

[3] A. Cooper, R. Reimann, D. Cronin, C. Noessel, About Face: The Essentials of Interaction Design, 4 ed., John Wiley & Sons, 2014.

[4] J. Pruitt, J. Grudin, Personas: Practice and theory, in: Proceedings of the 2003 Conference on Designing for User Experiences (DUX '03), Association for Computing Machinery, 2003, pp. 1–15. doi:10.1145/997078.997089.

[5] J. M. Carroll, Making Use: Scenario-Based Design of Human-Computer Interactions, MIT Press, 2000. doi:10.7551/mitpress/4398.001.0001.

[6] J. Kalbach, Mapping Experiences: A Complete Guide to Creating Value through Journeys, Blueprints, and Diagrams, O'Reilly Media, 2016.

[7] M. J. Bitner, A. L. Ostrom, F. N. Morgan, Service blueprinting: A practical technique for service innovation, California Management Review 50 (2008) 66–94. doi:10.2307/41166446.

[8] M. Obaidi, J. Klünder, Development and application of sentiment analysis tools in software engineering: A systematic literature review, in: Proceedings of the Conference on Evaluation and Assessment in Software Engineering (EASE '21), Association for Computing Machinery, 2021, pp. 80–89. doi:10.1145/3463274.3463328.

[9] M. Obaidi, L. Nagel, A. Specht, J. Klünder, Sentiment analysis tools in software engineering: A systematic mapping study, Information and Software Technology 151 (2022) 107018. doi:10.1016/j.infsof.2022.107018.

[10] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, R. Oliveto, Sentiment analysis for software engineering: How far can we go?, in: Proceedings of the 40th International Conference on Software Engineering (ICSE 2018), Association for Computing Machinery, 2018, pp. 94–104. doi:10.1145/3180155.3180195.

[11] N. Novielli, D. Girardi, F. Lanubile, A benchmark study on sentiment analysis for software engineering research, in: Proceedings of the 15th International Conference on Mining Software Repositories (MSR '18), Association for Computing Machinery, 2018, pp. 364–375. doi:10.1145/3196398.3196403.

[12] E. Guzman, W. Maalej, How do users like this feature? a fine grained sentiment analysis of app reviews, in: 2014 IEEE 22nd International Requirements Engineering Conference (RE 2014), IEEE, 2014, pp. 153–162. doi:10.1109/RE.2014.6912257.

[13] L. Zhao, W. Alhoshan, A. Ferrari, K. J. Letsholo, M. A. Ajagbe, E. V. Chioasca, R. T. Batista-Navarro, Natural language processing for requirements engineering, ACM Computing Surveys 54 (2021) 3444689. doi:10.1145/3444689.

[14] J. Cleland-Huang, O. C. Z. Gotel, J. Huffman Hayes, P. Mäder, A. Zisman, Software traceability: Trends and future directions, in: Proceedings of the on Future of Software Engineering (FOSE 2014), Association for Computing Machinery, 2014, pp. 55–69. doi:10.1145/2593882.2593891.

[15] Y. N. Tomoyuki Kajiwara, Wrime: Dataset for emotional intensity estimation, 2022. URL: https://github.com/ids-cv/wrime.

[16] T. N. Group, Bert base japanese (character-level tokenization with whole word masking, cc-100 and jawiki-20230102), 2023. URL: https://huggingface.co/tohoku-nlp/bert-base-japanese-char-v3.