Replication in Requirements Engineering: The NLP for RE Case

SALLAM ABUALHAIJA, University of Luxembourg, Luxembourg, Luxembourg
F. BASAK AYDEMIR, Utrecht University, Utrecht, Netherlands
FABIANO DALPIAZ, Utrecht University, Utrecht, Netherlands
DAVIDE DELL'ANNA, Utrecht University, Utrecht, Netherlands
ALESSIO FERRARI, CNR ISTI, Pisa, Italy
XAVIER FRANCH, Universitat Politécnica de Catalunya, Barcelona, Spain
DAVIDE FUCCI, Blekinge Institute of Technology, Karlskrona, Sweden

Natural language processing (NLP) techniques have been widely applied in the requirements engineering (RE) field to support tasks such as classification and ambiguity detection. Despite its empirical vocation, RE research has given limited attention to replication of NLP for RE studies. Replication is hampered by several factors, including the context specificity of the studies, the heterogeneity of the tasks involving NLP, the tasks' inherent *hairiness*, and, in turn, the heterogeneous reporting structure. To address these issues, we propose a new artifact, referred to as ID-CARD, whose goal is to provide a structured summary of research papers emphasizing replication-relevant information. We construct the ID-CARD through a structured, iterative process based on design science. In this article: (i) we report on hands-on experiences of replication; (ii) we review the state-of-the-art and extract replication-relevant information: (iii) we identify, through focus groups, challenges across two typical dimensions of replication: data annotation and tool reconstruction; and (iv) we present the concept and structure of the ID-CARD to mitigate the identified challenges. This study aims to create awareness of replication in NLP for RE. We propose an ID-CARD that is intended to foster study replication but can also be used in other contexts, e.g., for educational purposes.

CCS Concepts: • **Software and its engineering** → **Reusability**; **Requirements analysis**; • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Requirements Engineering (RE), Natural Language Processing (NLP), replication, tool reconstruction, annotation, ID card

1 INTRODUCTION

System requirements are primarily written in **natural language (NL)** [24, 27]. To analyze and manage textual requirements, the **requirements engineering (RE)** community has long been interested in applying **natural language processing (NLP)** technologies. The emerging research strand, **natural language processing for requirements engineering (NLP4RE)**, has received a lot of attention from both industry practitioners and academic researchers, leading to initiatives such as the NLP4RE workshop series [15]. To process textual information, NLP4RE tools utilize NLP techniques, which, in turn, rely on a variety of algorithms dominated by **machine learning (ML)**, **deep learning (DL)**, and **large language models (LLMs)**. A recent systematic mapping study [57] reports that the majority of research papers in NLP4RE (≈84%) involve proposing novel solutions or validating existing technologies. However, only a small fraction of developed tools (≈10%) is made publicly available. Unavailable artifacts can impede the positioning of novel solutions through sound comparisons against existing ones and further hamper the industrial adoption of NLP4RE tools.

Replication is an important aspect of empirical evaluation that involves repeating an experiment under similar conditions using a different subject population [50, 54]. Replicability is currently regarded as a major quality attribute in **software engineering (SE)** research, and it is one of the main pillars of Open Science [42]. The ACM badge system was introduced at the end of the 2010s[1] to reward, among others, available and replicable research. Several major conferences—e.g., ICSE, ESEC/FSE, and RE—feature an Artifact Evaluation Track, which grants variants of the ACM badges. Two mapping studies, from Da Silva et al. [12] and from Cruz et al. [11], cover replications in SE from 1994 to 2010, and from 2013 to 2018, respectively. In both studies, the RE field appears to be among the ones in which replication is most common. However, with a few exceptions (e.g., [13]), replication does not seem commonly practiced in NLP4RE. This lack of replication can be attributed to various challenges, with one of the most prominent being the incomplete reporting of studies, as noted by Shepperd et al. [49].

In this article, we propose a new artifact, referred to as ID-CARD, that fosters the replication of NLP4RE studies. The ID-CARD is a template composed of 47 questions concerning replication-relevant information, divided into seven topics. These topics characterize: the RE task addressed in the study; the NLP task(s) used to support the RE task; information about raw data, labeled datasets and annotation process; implementation details; and information related to the evaluation of the proposed solution. We advocate attaching the ID-CARD as part of the submission for future NLP4RE papers as well as creating it in retrospect for existing papers. The ID-CARD can be created and/or used by authors, newcomers to the field, reviewers, and students.

Defining the ID-CARD is triggered by our hands-on experience concerning two replication scenarios. The first scenario involves replicating a state-of-the-art baseline [55] as part of building a benchmark for handling anaphoric ambiguity in textual requirements. The second scenario involves replicating a widely used solution for classifying requirements into functional or non-functional [38] against which we compare and position a novel solution proposed by a subset of the authors of this article, previously published in [13]. Replication can be *exact* when one follows

---

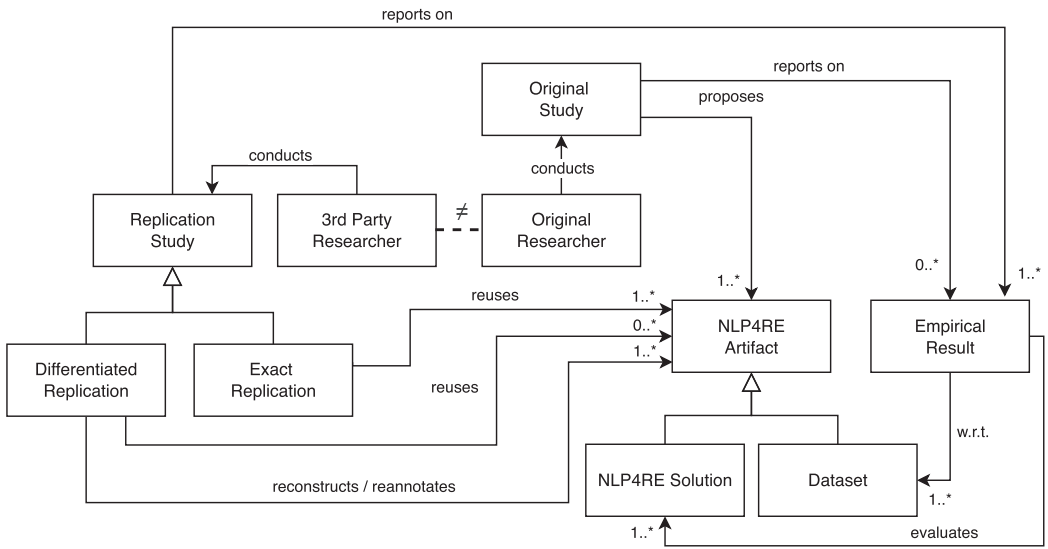[1]https://www.acm.org/publications/policies/artifact-review-and-badging-current. Visited April 5th, 2024.

Fig. 1. Overview of the terminology used in this articl.

the original procedure as closely as possible or *differentiated* when one adjusts the experimental procedures to fit the replication context [34]. In our replication experience, exact replication was not possible mainly due to the incomplete reporting of implementation details in the original papers. Deciding on such unknown details during replication can alter the original study to some extent. Minimizing decisions due to unknown details is an essential motivation for the ID-CARD. Building on our hands-on experience, we conducted several *focus groups* through which we identified the different challenges that can be encountered with regard to extracting replication-relevant information from NLP4RE studies. To address these challenges, we defined an initial version of the ID-CARD. We then reviewed a representative sample from the NLP4RE literature and iteratively refined the ID-CARD according to our findings and observations.

Terminology. Figure 1 provides an overview of the terminology used in this article. *Replication* refers to the attempt, conducted by *third-party researchers* (different from the *original researchers*), to obtain the same results of a specific *original study. Reuse* can be a prerequisite for replication: for replicating a study, third-party researchers may reuse the *NLP4RE artifacts* (*NLP4RE solutions* or *datasets*) that were proposed and released by the original researchers. Note that the figure focuses on reuse in the context of replication, although researchers may also reuse artifacts for other purposes. We refer to the information that can be extracted from the original paper for the purpose of replication as *replication-relevant information*. We further use *solution* and *tool* interchangeably to indicate the automated procedure described in an NLP4RE paper to solve a particular problem. We use the term *reconstruction* to denote the (re-)implementation of a solution as explained in the original paper.

Contributions. This article makes the following contributions:

(1) We identify a total of 16 challenges that can arise in practice during the replication of NLP4RE studies. We do so by relying on our hands-on experience in which we have reconstructed two NLP4RE solutions in addition to performing an in-depth review of existing papers spanning diverse topics in the NLP4RE literature. As replication in NLP4RE often requires creating a dataset, we differentiate between the challenges related to dataset *(re-)annotation*—aimed to determine the gold standard against which the solution is

evaluated—versus challenges concerning tool *reconstruction*. The final list of challenges is presented in Section 5.

(2) We devise an ID-CARD that summarizes in a structured way the replication-relevant information in NLP4RE papers. We demonstrate the applicability of the ID-CARD by manually creating the equivalent ID-CARDs for a total of 46 research papers from the NLP4RE literature. Specifically, we manually extracted the replication-relevant information from these papers and provided the answers to the questions posed in the ID-CARD. For 15 out of the 46 papers, we also let the original authors independently fill in the ID-CARD. We also let the authors assess several aspects of the proposed ID-CARD, such as its ease of use. As we discuss in Section 6, the results indicate that the ID-CARD can be used as a complementary source to the original paper for facilitating replication. We make all the material publicly available online [1].

Structure. Section 2 discusses related work about replicabilities and associated challenges. Section 3 discusses the research method and research questions. Section 4 provides the context for the two replication scenarios considered in our work. Section 5 describes the challenges of annotation and tool reconstruction. Section 6 presents the ID-CARD. Section 7 discusses issues related to the use of the ID-CARD and the limitations of the study. Finally, Section 8 concludes the article and sketches future directions.

2 RELATED WORK

This article focuses on replication in NLP4RE, a sub-field of RE that belongs to the broad area of SE. Only a few replication studies are published in this field (e.g., [20, 48]), showing the need for methods and tools that can facilitate conducting such studies. Other studies in related domains have also identified some of the issues discussed in our article [2, 46, 58], although those works build on secondary research (they are *surveys*), while our approach heavily relies on our first-hand experience. In the following, we first discuss initiatives in Open Science and replicability in SE to show the relevance of this topic in the current discourse of the SE community and the complexity of ensuring replicability. Then, we refer to relevant works discussing guidelines and issues of study replication to illustrate the addressed research gap.

Open Science and Replicability Initiatives. Open Science in SE is a movement aiming to make research artifacts—including source code, datasets, scripts to analyze the data, and manuscripts—openly available to the research community [42]. This enables *replicability*, i.e., the possibility for other research teams to repeat a study by reusing the artifacts provided by the authors of the original work or by reconstructing them. Moreover, Open Science facilitates verifiability and transparency, which are nowadays critical research drivers for SE in general and RE in particular. For example, ICSE 2024 review criteria define verifiability and transparency as "The extent to which the paper includes sufficient information to understand how an innovation works; to understand how data was obtained, analyzed, and interpreted; and how the paper supports independent verification or *replication* of the paper's claimed contributions".[2] Major SE outlets have introduced Open Science Policies (e.g., ESEC/FSE 2024[3]), requiring authors to make their artifacts available to the reviewers so that research results can be scrutinized during the review phase and made reproducible once published. Moreover, conferences include Artifact Evaluation Tracks aimed at explicitly reviewing artifacts and assigning badges (e.g., the ones defined by ACM[4]) to the

---

[2]https://conf.researchr.org/track/icse-2024/icse-2024-research-track#Call-for-papers. Visited January 4th, 2024.
[3]https://2024.esec-fse.org/track/fse-2024-open-science-policy. Visited January 4th, 2024.
[4]https://www.acm.org/publications/policies/artifact-review-and-badging-current. Visited January 4th, 2024.

associated papers. These tracks typically include a separate **Program Committee (PC)** highlighting not only the relevance given to Open Science but also the *complexity* of replicability assessment, which thus requires a dedicated PC. The guidelines for artifact evaluation became quite extensive, as shown, for example, by the ICSE 2021 guidelines,[5] which count 13 pages, as the evaluation process needs to ensure the fulfillment of multiple fine-grained aspects. To account for the complexity of replicability and the early degree of maturity of the community, the PC members are typically required to interact with the authors to provide them feedback to fix the artifacts before these can be evaluated—this is usually the case of software or scripts that cannot be properly run in order to produce the expected results. In addition to Artifact Evaluation Tracks, the ROSE festival (Rewarding Open Science Replication and Reproduction in SE[6]) has been held in a number of SE top venues. The festival includes enlightening talks about replication studies to raise awareness in the community. In this scenario, not only conferences but also journals are also catching up with Open Science. The *Empirical Software Engineering Journal (EMSE)* introduced the OSI Open-Data badge [43] and *ACM Transactions on Software Engineering Methodology (TOSEM)* encourages **Replicated Computational Results (RCR)** Reports that should complement papers with information to support replications.[7] This landscape of initiatives demonstrates the relevance given to replicability in SE, and the complexity of ensuring verifiability and transparency.

Replicability Issues, Guidelines, and Challenges. From an online questionnaire of the *Nature* journal involving 1,576 researchers from different disciplines, it emerged that 70% of the respondents tried and failed to reproduce another scientist's experiment [4], indicating that replication is an issue for the wider research community. The main causes listed (60% of the respondents) are pressure to publish and selective (i.e., incomplete) reporting. Furthermore, the absence of methods or code, as well as raw data, are regarded as relevant by more than 40% of the participants. Among the solutions, the authors propose more academic rewards for replicability-enabling content, as well as journal checklists to ensure that replication-relevant material is included. Concerning the specific field of SE, multiple mapping studies and survey papers have been published on Open Science and replication, identifying the current status, guidelines, and open challenges. Da Silva et al. [12] cover replications in SE from 1994 to 2010. The study highlights that replications have substantially increased in the period considered and discusses desired characteristics of the replication studies. The study also includes a useful quality checklist to evaluate replication studies based on Carver's guidelines [7]. De Magalhães et al. [19] cover replications from 1994 to 2013. The authors list recommendations oriented to those who perform replications, but also conditions that must be fulfilled by the original papers to be replicated. The latter include the need to provide all necessary replication-relevant details, precise and unambiguous definitions, and the sharing of raw data and tool versions. Gonzalez-Barahona et al. [30, 31] evaluate the replicability[8] of **Mining Software Repositories (MSR)** studies, and propose a set of best practices to make a study replicable and applicable beyond the MSR field. These are mainly focused on features of the replication package, which should include raw data, processed data, results, links to reused data

---

[5]https://conf.researchr.org/getImage/icse-2021/orig/Submission.and.Reviewing.GuidelinesNUEVO.pdf. Visited January 4th, 2024.

[6]https://github.com/researchart/rose/tree/master. Visited January 4th, 2024.

[7]https://dl.acm.org/journal/tosem/replicated-computational-results. Visited January 4th, 2024.

[8]The authors use the term *reproducibility* rather than *replicability*, but their definition of the term is similar to our definition of replicability: "The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts". For this reason, we use the term *replicability* to avoid confusion.

(if any), and source code. In the context of the NLP4RE, a secondary study by Ahmed et al. [2] shows that published results in the area of requirements formalization from natural language are often non-reproducible due to lack of access to tools, data, and critical information not reported in the primary studies. Similarly, another systematic review in the same area [46] shows that the lack of openly accessible benchmarks hinders requirement formalization research.

In a book chapter, Mendez et al. [42] characterize Open Science from multiple perspectives, including sharing manuscript preprints and artifacts, and provide detailed guidelines. The chapter is among the few works that discuss challenges related to Open Science. These include the overhead of sharing data, privacy, confidentiality, licensing, appropriate preparation of qualitative data, and anonymity issues—particularly in sharing preprints to be submitted to venues that adopt the double-blind review model. These are mainly challenges from the perspective of the authors of the original study to be replicated rather than from the third-party researchers who want to perform a replication. Finally, Anchundia and Fonseca [3] provide guidelines that can facilitate replications in SE and informally list challenges that prevent or hamper replications: "reports and packages are neither sufficient to capture all information (e.g., raw material) nor to share tacit knowledge", "large amount of effort to obtain all necessary data", "replications do not satisfy professional needs", "the cost of conducting replications".

Research Gap. The landscape of Open Science initiatives in SE shows the increasing interest of the SE community in Open Science in general and replicability in particular underscoring that replication is a complex task, which requires further investigation and appropriate tool support. Compared with previous studies, we notice that a large part of them list criteria to be fulfilled to enable replication [12, 19, 30, 31], but only two of them [3, 42] identify *general* replication-related challenges. None reports a *context-specific* list of issues, specifically considering the viewpoint of third-party researchers who want to replicate a study, as in our case. While some general challenges also apply to our case, the provided list targets the NLP4RE context, which makes the identified issues more concrete. Furthermore, compared with previous work, our main outcome is a practical means to summarize relevant information to enable replications—i.e., the ID-CARD, which aims to address the identified challenges.

3   RESEARCH QUESTIONS AND METHOD

The research goal of this article is to support the extraction and documentation of replication-relevant information from NLP4RE papers. To this end, we propose two main dimensions that characterize replication in the context of NLP4RE: (i) the *datasets* with their annotations, as research in NLP inherently relies on datasets for evaluation as well as for developing diverse solutions, e.g., training an ML classifier; and (ii) the reconstruction of the proposed *tools*, as most NLP4RE studies describe an automated NLP-based solution that tackles an RE problem [57], which typically needs to be reconstructed.

To achieve this goal, we devise a set of **research questions (RQs)**. The first two questions are instrumental to responding to the third one:

RQ1:   *What are the challenges of annotating datasets for training and evaluating NLP4RE tools?*
As stressed in previous studies [15, 25, 57], the number of *annotated* (i.e., labeled) datasets in NLP4RE is scarce. Concrete annotation guidelines are essential to ensure the soundness of both reuse and replication. Reusing an existing (available) tool on different datasets (e.g., based on industrial data) requires annotating these datasets following the same annotation procedure as the original dataset. Replicating the entire study also requires applying the same annotation procedure. Missing guidelines can cause unwanted differences in results

between the original study and the replicated one. In RQ1, we investigate the challenges of creating labeled datasets while making them available to and reusable by the research community. We also consider the case in which an existing dataset must be re-annotated for various reasons to support replication.

RQ2: *What are the challenges in reconstructing NLP4RE tools?*

In RQ2, we study the challenges that third-party researchers encounter when reusing and/or replicating existing tools as part of their NLP4RE research. Since the majority of NLP4RE tools are unavailable [57], we highlight the process of *reconstructing* (i.e., developing) a tool using only the descriptions of the algorithms provided in the research papers in which the tool is presented. More concretely, the notion of reconstruction in our context refers to the cases in which the solution is exclusively described in the research paper yet is not complemented with any source code. While such cases can be regarded as extreme, they currently represent the common situation of tool reuse specifically in NLP4RE. Dealing with the extreme cases encapsulates as well the challenges that are pertinent to less challenging replication scenarios.

RQ3: *How to support NLP4RE researchers to overcome the identified challenges in RQ1 and RQ2?*

In response to RQ3, we propose developing an artifact that complements the NLP4RE research papers to primarily facilitate replication. The artifact, which we refer to as ID-CARD, is presented in Section 6.

To answer RQ1 to RQ3, we follow the research method sketched in Figure 2. The figure also serves as a reading guide for the various sections of this article. To facilitate reading, the method is presented here linearly. However, the process was conducted iteratively.

In order to respond to RQ1 and RQ2, we reconstructed two established NLP4RE tools for which the source codes were not available (see Step ① in Figure 2) [38, 55]. The reconstruction naturally required the (re-)annotation of datasets. The first tool, originally proposed by Yang et al. in 2011 [55], concerns anaphoric ambiguity handling, a long-standing, highly researched problem in NLP4RE [57]. The second tool, originally proposed by Kurtanović and Maalej in 2017 [38], concerns functional and non-functional requirements classification, a very popular RE task that was also selected for the *dataset challenge* at the IEEE Requirements Engineering conference in 2017 (RE'17). These two tools were reconstructed by two subgroups of our research team at different times for different purposes, as we elaborate in Section 4.

The rest of the process was mainly driven by focus groups conducted according to the guidelines from Breen [6]. Applying this research instrument, the cases and the collected data act as a trigger for reflection so that the participants can brainstorm about specific challenges they encountered, recall previously faced problems, and compare their viewpoints with their peers. We refer to Appendix A for further details on how the focus groups were conducted (structure, participants, timing, analysis, etc.).

The first two focus groups (Step ② and Step ③ in Figure 2) led to identifying a set of challenges related to RQ1 and RQ2. The challenges were reflected during the development of the NLP4RE ID-CARD (Step ④).

Step ④ included iterative activities consisting of multiple rounds of design and assessment among the authors. The activity was supported by an analysis of the literature (Step ⑤ in Figure 2) in which we read and analyzed 46 representative NLP4RE papers, selected based on the literature review by Zhao et al. [57]. A pair of researchers then used a similar strategy as the one applied in the mapping study to retrieve 155 additional papers published after 2019—i.e., the date of the mapping study. Finally, we selected 46 papers considering the following criteria: (1) Ensuring a balanced set of highly cited papers (i.e., representing the state-of-the-art) as well as more
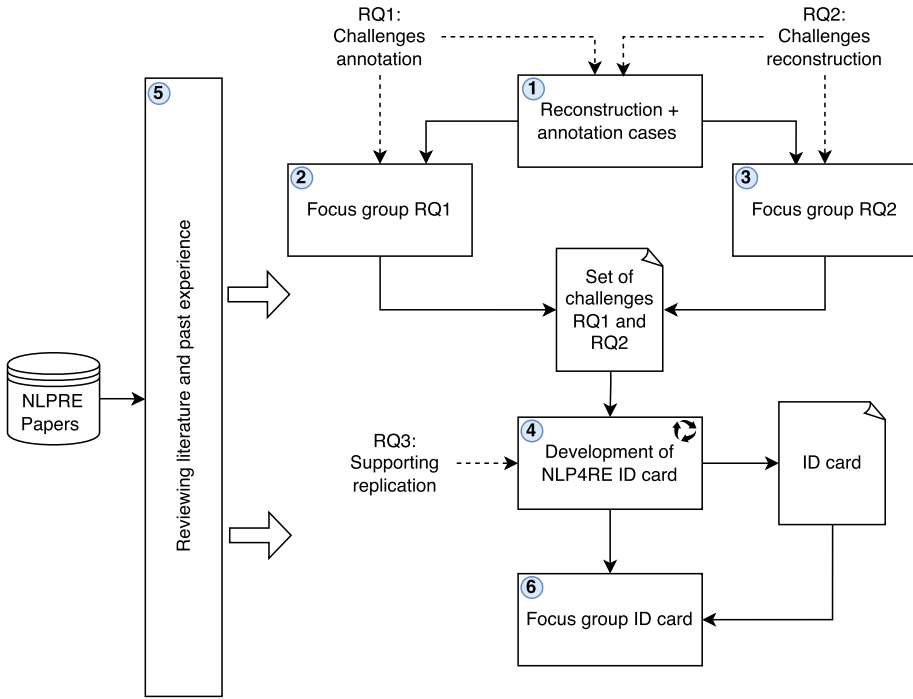
Fig. 2. Overview of the research method of this article.

recent papers. (2) Inspired by the topic categorization of the NLP4RE landscape [57], our selected papers covered the following nine categories: requirements classification, tracing, defect detection, model generation, test generation, requirements retrieval, information extraction from both legal documents and requirements documents, and app review analysis. We selected five papers from each category. At least two of the authors of this article independently manually analyzed a subset of the research papers and then the findings were cross-checked (*internal assessment*). The objective was to assess the completeness, cohesiveness, and coverage of the ID-CARD. In the final design iteration, we shared the NLP4RE ID-CARD with 15 authors of the NLP4RE papers that we considered in our design. We let them fill in the ID-CARD for their papers and further evaluate their own experience by answering a questionnaire that we appended to the ID-CARD, as we explain in Section 6.4 (*external assessment*). To conclude, we organized a third focus group (Step ⑥) in which we discussed the ID-CARD and its use, leading to a number of observations to which we refer in the discussion (Section 7).

The research method of Figure 2 aligns with Wieringa's design science research, particularly with the design cycle [51], which comprises problem investigation, treatment design, and treatment validation. In particular, *problem investigation* is conducted in Steps ① through ③, wherein the researchers conduct reconstruction and annotation cases. These activities, complemented by the analysis of the literature conducted in Step ⑤ and consolidated by two focus groups, led to the identification of a set of challenges for RQ1 and RQ2. These challenges inform the *treatment design*—the NLP4RE ID-CARD (Step ④). *Treatment validation* of the ID-CARD has been conducted both via the internal and external assessment activities.

In essence, multiple iterations of the design cycles were conducted as Step ④ has a cyclic nature in Figure 2. The last validation activity was conducted via a focus group (Step ⑥) in which the ID-CARD was finalized to its current form.

4 RECONSTRUCTION CASES OF NLP4RE SOLUTIONS

We share our hands-on experience in replicating two state-of-the-art solutions from the NLP4RE literature (Step① in Figure 2). The first solution (Section 4.1) focuses on detecting anaphoric ambiguity, a specific type of defect in NL requirements, as presented in an early work by Yang et al. [55]. The second solution (Section 4.2), presented in a more recent work by Kurtanović and Maalej [38], tackles the classification of non-functional requirements. Both defect detection and requirements classification are among the most common tasks in NLP4RE field [57]. These topics and associated papers have garnered substantial recognition, with both papers accumulating over 160 citations according to Google Scholar,[9] making them highly representative of the NLP4RE field.

4.1 Anaphoric Ambiguity Detection

Motivation and overview. Ambiguity in NL requirements is a long-standing research topic in RE [5, 16, 23, 28, 35, 37]. Pronominal anaphoric ambiguity occurs when a pronoun can refer to multiple preceding noun phrases. The work of Yang et al. [55] was chosen as a representative approach in anaphoric ambiguity detection. Since neither the dataset nor the source code was to be found publicly, we decided to reconstruct the work using the details in the original paper and adapt it to the needs of our contest. This annotation activity acted as a trigger for the focus group in Step ②. The RE literature discusses unconscious disambiguation, in which the stakeholders involved in a given project are able to disambiguate the requirements thanks to their domain knowledge [18, 22, 45, 47]. Automated detection and resolution of different ambiguities is still a valid scenario in RE since not all stakeholders have the same level of domain knowledge and would hence resolve incorrectly. Another scenario is end-to-end automation needing to be developed for other purposes, e.g., extracting information from requirements might necessitate resolving referential ambiguity as a prerequisite.

Annotation and Dataset Creation. We used the **PURE (PUblic REquirements)** dataset [26] and randomly selected a subset of 200 requirement statements from seven domains, such as railway and aerospace. Each requirement statement constituted one sentence, typically using the 'shall' format. The rationale behind our selection was that the resulting set should be manageable, time- and effort-wise, for annotation within the time frame of our contest. Further, covering diverse domains could be advantageous to assess the performance of a given solution across domains.

Prior to annotating the requirements, we exchanged some guidelines about the ambiguity task. The task was to decide whether a pronoun occurrence in a given requirement is ambiguous or not by investigating the relevant antecedents. Each pronoun occurrence was analyzed by two annotators from the participants. Our annotation process was then performed in multiple rounds. After each round, we had a session to discuss our findings and some problematic issues. The annotation process resulted in 103 ambiguous requirements (i.e., containing an ambiguous pronoun occurrence). For all remaining requirements marked as unambiguous, the annotators were asked to identify the antecedent they deemed correct. The outcome of our annotation process is a dataset in which each pronoun occurrence is labeled as ambiguous if it is (i) marked as such by at least one annotator or (ii) interpreted differently by the two annotators. Otherwise, when two annotators agree on the same interpretation, the pronoun occurrence is labeled unambiguous. We computed the pairwise inter-rater agreement using Cohen's Kappa [40] on a subset of ≈8% from our dataset (16 requirements randomly selected). We obtained an average Kappa of 0.69, suggesting "substantial agreement" between the annotators. Following common practices, we used disagreements as indicators of ambiguity.

---

[9]Citations were collected on January 12, 2024.

Tool Reconstruction. The solution proposed by Yang et al. [55, 56] combines ML and NLP technologies to detect anaphoric ambiguity in requirements, taking a set of textual requirements as input and determining whether each pronoun occurrence is ambiguous or not. We recently reconstructed this solution and publicly released it [22].

The reconstructed version of the approach includes four components: text preprocessing, pronoun–antecedent pair identification, classification, and anaphoric ambiguity detection. The first component (*text preprocessing*) parses the textual content of the requirements document and applies an NLP pipeline consisting of four modules: (i) a tokenizer for separating out words from the running text, (ii) a sentence splitter for breaking up the text into separate sentences, (iii) a **part-of-speech (POS)** tagger for assigning a part of speech to each word in the text (e.g., noun, verb, and pronoun), and (iv) a chunker (or a constituency parser) to delineate phrase boundaries, e.g., **noun phrases (NPs)**.

The second component (*pronoun–antecedent pair identification*) extracts all pronouns occurring from the input requirements, identifies a set of likely antecedents for each pronoun, and finally generates a set of pronoun–antecedent pairs. For the purpose of simplifying our reconstruction scenario and since coreference resolution was not directly relevant to our contest, we left it out.

The third component (*classification*) builds an ML-based classifier to classify a given pair of a pronoun p and a relevant antecedent a into *YES* when p refers to a, *NO* when p does not refer to a, or *QUESTIONABLE* when it is unclear whether p refers to a. The classifier is trained over a set of manually crafted language features that characterize the relation between p and a, e.g., whether p and a agree in number or gender. The original paper lists 17 features divided in three categories: 11 syntactic and semantic features, two document-based, and three corpus-based features. We dropped out four features that are computed using a proprietary library and related to sequential and structural information to facilitate the reusability of our reconstructed solution.

The last component (*ambiguity detection*) applies a set of rules over the predictions produced by the ML-based classifier to distinguish the ambiguous cases. The original paper reports two thresholds for detecting correct and incorrect antecedents for unambiguous cases. In our reconstructed solution, we redefined these thresholds with empirically optimized values for our dataset. The thresholds in the reconstructed tool could be generalized beyond the dataset as supported by empirical evidence in the paper, which reports the reconstruction [22]. The reconstructed code is available in an online repository [21].

4.2 Functional and Non-functional Requirements Classification

Motivation and Overview. Motivated by the importance of identifying quality aspects in a requirements specification starting from the early stages of SE, ML and NLP techniques have been widely used in proposing solutions to several related problems [44], such as classifying requirements into functional vs. non-functional [10], which we refer to as the *FR-NFR classification problem*. Despite the variety of existing FR-NFR classification problem approaches, as of 2019, the most effective reported classifiers [38] relied on a characterization of the requirements at the word level, e.g., via text n-grams or POS n-grams, resulting in a large number of low-level features (100 or 500 in [38]). Moreover, for evaluating the classifiers, the great majority of the literature in requirements classification focuseson the PROMISE NFR dataset [9], a collection of 625 requirements from 15 projects created and classified by graduate students.

Aiming at a more interpretable solution than those relying on a large number of low-level (word-level) features, which make it hard for analysts to understand why the classifier performs well or poorly and why requirements are classified in a certain way, in 2019, a subset of authors of this article proposed a new NLP approach to the FR-NFR classification problem [13]. In proposing and

evaluating the new solution, we (i) manually annotated a set of 1,500+ requirements consisting of 8 different projects and (ii) reconstructed Kurtanović and Maalej's tool [38] to compare the proposed solution with the state-of-the-art.

Dataset and Annotation. We manually re-annotated 1,500+ requirements from the PROMISE dataset [9] and 7 industrial projects. For the annotation, we followed an approach based on the taxonomy of Li et al. [41], which allows a requirement to possess both functional and non-functional aspects, as opposed to the original annotation, which allowed a single label per requirement. In particular, we annotated a requirement as possessing functional aspects (F) if it included either a functional goal or a functional constraint, whereas we annotated a requirement as possessing quality aspects (Q) (and thus being a non-functional requirement) if it included a quality goal or a quality constraint. The decision on the functional aspect was independent of the decision on the quality aspect. Thus, a requirement could possess only F aspects, only Q aspects, both aspects (F+Q), or none. In the last case, we considered the requirement as denoting auxiliary information [52].

Each dataset was independently manually analyzed by two annotators. Reconciliation meetings were then organized to review the disputes in the annotation. If the annotators failed to convince each other, a third annotator was consulted for the final label. The annotators went over all disagreements and managed to resolve them.

Tool Reconstruction. To provide insights about how our approach compared against the state-of-the-art, we selected a relatively recent approach by Kurtanović and Maalej [38], described extensively in the original paper, which shows excellent performance. The original solution consists of characterizing the training (PROMISE NFR) dataset using several low-level word features, such as n-grams, or POS n-grams. The authors of the original publication consider two cases with the top (most discriminant) 100 and 500 features. The requirements, labeled with the top features, are then used to train a **Support Vector Machine (SVM)** model with a linear kernel and use the trained model to classify previously unseen requirements as functional or non-functional.

Since the original classifier was not publicly available, we reconstructed it from the details provided in the original publication. We also complemented such information with a code of another classifier related to app reviews [39] that is partially available online and developed by the same research group.

During reconstruction, we applied a few minor modifications to the original version (details in [13]). For example, to build parse trees, we used a different better-performing library (Berkeley's benepar [36]) than that used in the original publication (the Stanford parser [8]). We could not reproduce one of the classifier's features since the explanation in the original paper was insufficient for us to make a correct re-implementation. We could not use the same dataset applied in the original solution to artificially balance the minority class of NFRs, as it was not publicly released. Finally, we released the reconstructed code in an online repository [14].

5   REPLICATION CHALLENGES OF NLP4RE STUDIES

In this section, we present a total of 16 challenges derived from our first two focus groups complemented by our hands-on experience. Table 1 lists the identified challenges in relation to dataset annotation (RQ1, Step ② in Figure 2) and tool reconstruction (RQ2, Step ③).

5.1   Challenges of Dataset Annotation and Re-annotation (RQ1)

During our focus group for RQ1, we identified 10 challenges pertinent to creating a labeled dataset for a specific RE task. We organize the challenges into four categories: (a) *theoretical foundation* (Ann1 and Ann2), (b) *annotation process* (Ann3 – Ann5), (c) *dataset-related* (Ann6 – Ann8), and (d) *human and social aspects–related* (Ann9 and Ann10).

Table 1. Summary of Challenges Pertinent to Replication of NLP4RE Studies

(a) Dataset Annotation Challenges (**RQ1**)

ID	Description
Ann1	Some RE-specific categorization tasks lack solid theories that can guide the annotation process.
Ann2	Besides annotation experience and theoretical knowledge, the lack of domain knowledge can limit the accuracy of the annotations.
Ann3	The annotation activity is time-consuming due to factors such as language barriers, different individuals' backgrounds, and fatigue.
Ann4	The annotation protocol can evolve and thereby necessitate the re-annotation of the data, which might, again, cause additional time and effort.
Ann5	Theoretical and practical training resources and opportunities are limited and not adequate for training novice annotators who are often trained during the annotation task by more experienced annotators.
Ann6	The lack of benchmarks entails that annotated datasets enabling comparison against the state-of-the-art are scarce.
Ann7	Available imbalanced datasets pose the challenge of both understanding the minority class and consequently the annotation of new examples thereof.
Ann8	Determining the right amount of context to be shared along with the annotation raw data is essential and can significantly affect the annotation results.
Ann9	Motivating the annotators poses another challenge since an immediate observation of the impact of a given annotation task is not always possible.
Ann10	Annotators are often not experienced in managing the social aspects or resolving conflicts originating from power, authority, or other social relations.

(b) Tool Reconstruction Challenges (**RQ2**)

ID	Description
Rec1	The reconstruction-relevant information and implementation details of the original approach can be ambiguous, imprecise, and incomplete.
Rec2	If a tool was partially or fully developed and/or evaluated using proprietary data, then there is no guarantee that the reconstructed tool would be identical to the original one since the used data cannot be accessed for reconstruction purposes.
Rec3	Communication with the original authors is not always useful since the actual information sources may no longer be available.
Rec4	The continuous evolution of the NLP ecosystem entails that some libraries become outdated, unavailable, or are no longer maintained.
Rec5	Tools are typically developed as prototypes and not maintained in the long term.
Rec6	Tool reconstruction is not (yet) valued as a self-standing research contribution; hence, researchers are discouraged from replicating tools over time.

5.1.1 Theoretical Foundation Challenges.

Ann1. Some RE-specific categorization tasks lack solid theories that can guide the annotation process. Unlike typical NLP tasks, which rely on solid theories from linguistics, RE categorization tasks often lack a common and agreed-upon theory defining the different classes. For example, thus far, researchers have proposed numerous definitions and taxonomies for classifying requirements into functional and non-functional categories [29, 41]. For instance, the requirement *"The system shall back up the customer data every 2 hours"* can be labeled as non-functional according to Glinz's definition [29], since it conveys a specific quality concern (i.e., reliability). Conversely, the exact requirement can be labeled as both functional and non-functional according to the definition by Li et al. [41], since it has a functional part (i.e., to back up customer data) as well as a quality that qualifies it in terms of performance (i.e., every 2 hours). Since neither of these two definitions prevails over the other, the annotation output can differ depending on the annotators' background.

Ann2. Besides annotation experience and theoretical knowledge, the lack of domain knowledge can limit the accuracy of the annotations. For example, the lack of knowledge

about regulations, standards, and data types when annotating a dataset can result in confusing compatibility with compliance requirements. Furthermore, the annotators may often use existing datasets from a particular domain (e.g., automotive) without necessarily having adequate experience in that domain or any interactions with domain experts or the original annotators who created the datasets. For example, not knowing whether an abbreviation such as STANAG 4609 refers to a standard or regulation leads to confusion for identifying compatibility and compliance requirements, respectively.

5.1.2 Annotation Process Challenges.

Ann3. The annotation activity is time-consuming due to factors such as language barriers, different individuals' backgrounds, and fatigue. An annotation task in NLP4RE often deals with manually analyzing several requirements artifacts and assigning labels as required by the respective task. Additional delays can be caused by several factors, including, but not limited to, the following. (i) *Language barriers:* Most requirements texts are written in English. Since English is not necessarily the mother tongue of those who write the requirements or those who read them, handling requirements in English may lead to grammatical errors in the text and/or different interpretations. (ii) *Reconciliation of multiple annotations for the same sample:* Individuals' backgrounds might be a barrier to reaching full agreement. Even when the different annotators see each other's perspectives, sometimes they still insist on their own. (iii) *Fatigue:* The amount of effort spent on the annotation task causes considerable fatigue, which can affect the performance of the annotators. Defining clear fatigue thresholds and fatigue-mitigation strategies require finding a balance between throughput and quality of the resulting annotations.

Ann4. The annotation protocol can evolve and thereby necessitate the re-annotation of the data which might, again, cause additional time and effort. To increase inter-rater agreement, the annotation activity is often conducted iteratively, holding regular consolidation meetings. This may result in changes in the annotation protocol, which may compromise the quality of the annotation. Any change would immediately raise the question about whether to re-annotate the previous annotations and thus improve the quality or continue without doing so due to time considerations.

Ann5. Theoretical and practical training resources and opportunities are limited and not adequate for training novice annotators who are often trained during the annotation task by more experienced annotators. In an ideal scenario, annotators should be trained prior to the annotation task, because (i) annotators are expected to meet a minimum qualification level (e.g., domain knowledge, motivation, availability, and seriousness); and (ii) annotators will be more attentive when they learn about certain issues, e.g., the inter-rater agreement. However, the appropriate level of detail and training resources are often difficult to determine in advance. An alternative is to prepare guidelines and/or have small iterations (pilots to train/test people). Written guidelines, despite being useful, can be ambiguous (in particular, as future references after the annotation has been conducted). Hence, the same results are unlikely to be obtained by different annotators even if the same written guidelines are provided.

5.1.3 Dataset-Related Challenges.

Ann6. The lack of benchmarks entails that annotated datasets enabling comparison against the state-of-the-art are scarce. For example, currently, the most widely adopted dataset for the FR-NFR classification problem is the PROMISE dataset. This is a benchmark with questionable quality, since it contains requirements that were written and classified by students. More

recent work in the NLP4RE literature has provided a re-annotated version of that dataset [13, 41], clearly showing some discrepancy with the original annotated dataset.

Ann7. Available imbalanced datasets pose the challenge of both understanding the minority class and consequently the annotation of new examples thereof. The different labels are often unevenly distributed among the data. Given that there are few examples that are associated with the minority label, annotators will have less examples to learn about that label. For instance, the distributions of functional and non-functional requirements in our replication study were quite diverse and often skewed across the different annotated datasets, leaving the annotators with doubts when it comes to the minority class.

Ann8. Determining the right amount of context to be shared with the raw data to be annotated is essential and can significantly affect the annotation results. We observed in both reconstruction cases that not knowing the wider context of a given requirement to be annotated can complicate the annotation process. In the case of referential ambiguity, the context of a requirement may clarify the ambiguity. Similarly, in the case of non-functional requirements classification, the location of the requirement in the document or the surrounding requirements may hint the intended class.

5.1.4 Human and Social Aspects–Related Challenges.

Ann9. Motivating the annotators poses another challenge since an immediate observation of the impact of a given annotation task is not always possible. Inviting annotators from industry is difficult and sometimes not possible. Annotation *per se* has no direct application (or interest) in industry since practitioners (domain experts in the case of annotation) do not need annotated data, but rather a working solution. In addition, merely annotation tasks are not always appreciated in the research community and are thus hard to publish. The latter observation is confirmed by the majority of the focus group participants, who regard annotation as an unfulfilling task.

Ann10. Annotators are often not experienced in managing the social aspects or resolving conflicts from power, authority, or other social relations. The human aspects of an annotation process (e.g., the role of persuasion) are typically not covered by the annotation protocols and training. However, the social aspect is often necessary both to improve the understanding of each other's perspectives and consequently reach agreements. For any annotation task, social challenges cover different aspects, including the decision to objectively account for persuasion, the best schema to decide on consensus (e.g., consensus seeking, majority votes, senior person makes the final decision), motivating the annotators, power and authority relations, and the impact of peer pressure.

5.2 Challenges of Tool Reconstruction (RQ2)

During our focus group for RQ2, we identified a total of five challenges pertinent to tool reconstruction concerning (a) *information availability* (Rec1 – Rec3), (b) *technological divergence* (Rec4 and Rec5), and (c) *motivation* (Rec6).

5.2.1 Information Availability Challenges.

Rec1. The reconstruction-relevant information and implementation details of the original approach can be ambiguous, imprecise, and incomplete. The original paper is normally written to communicate a study rather than with the goal of facilitating tool reconstruction. Therefore, details about the architecture are frequently left out. While omitting implementation details

can make a paper more readable, missing such details can complicate tool reconstruction. Examples of such implementation details include choice of parameters, seed values for the experiments, configuration options, detailed feature descriptions, and (the exact versions of) libraries. The availability of the source code is not always useful for reconstruction since the accompanying documentation is often unclear. For instance, one of the major obstacles we encountered in reconstructing the requirements classification tool was related to understanding the learning features for the ML classifier. The authors of the original tool applied 11 different *types* of learning features, including text n-grams, POS n-grams, the fractions of nouns, verbs, adjectives, adverbs in the requirement, and more. While some of these features are standard and can be easily re-implemented (e.g., the depth of the syntax tree), other features were unclear, and their descriptions in the original paper were inadequate for an exact re-implementation. An example of unclear feature concerns the unigrams of POS tags on the clause and phrase level.

Rec2. If a tool was partially or fully developed and/or evaluated using proprietary data, then there is no guarantee that the reconstructed tool would be identical to the original one since the used data cannot be accessed for reconstruction purposes. Private or unpublished datasets can sometimes be used in developing (some steps of) the original tool or during validation. For example, in the considered FR-NFR classification problem, the original tool applied, in addition to the public PROMISE dataset, another dataset from Amazon software reviews to balance the minority class. This second dataset was also used for selecting the most relevant learning features. However, as we did not have access to this additional dataset, we had to resort to a different solution for balancing the minority class, compromising the selection of the same relevant features. Since the evaluation set also contained proprietary data, verifying the correctness of the reconstructed tool was not possible in this case.

Rec3. Communication with the original authors is not always useful since the actual information sources may no longer be available. It is fundamental to communicate with the authors of the original tool for assistance with the reconstruction. To this end, in some cases, limited responsiveness can be observed, in particular when the authors are contacted about very specific details (e.g., explanation of learning features). Ideally, the authors of the original tool should help validate the correctness of the reconstructed tool. However, in practice, the developers of a tool are often PhD or MSc students, who may have left the organization, taking away in-depth knowledge about the tool.

5.2.2 Technological Divergence Challenges.

Rec4. The continuous evolution of the NLP ecosystem entails that some libraries become outdated, unavailable, or are no longer maintained. An existing approach from the NLP4RE research often contains text preprocessing steps based on available NLP technologies (e.g., stopwords removal or lemmatization). Outdated NLP libraries do not always have equivalent replacements. Based on our experience, older libraries were applied, for example, in the case of requirements classification to construct syntax trees or undersample the majority class. Some of these old libraries became deprecated, incompatible with new versions of libraries doing other tasks, or performed worse than their respective new versions. Updating libraries can therefore introduce a discrepancy between what is expected and what is achieved, for better or worse. For instance, to create the parse trees accurately in one of our reconstructions, we opted for Berkeley's benepar library instead of the Stanford parser used in the original paper. This challenge was more notable during the reconstruction of the ambiguity handling tool that was proposed in 2011. Since then, the NLP technologies have been through several breakthroughs, changing the landscape for textual preprocessing.

Rec5. Tools are typically developed as prototypes and not maintained in the long term.
Developed NLP4RE tools are often left in their initial prototype status and not maintained in the
long term. As a consequence, tools become harder to reuse over time: the availability of platforms
and operating systems changes over time and specific supporting software (which itself may no
longer be available) can block reconstruction. One of the reasons emerging from our focus groups
for lack of maintenance is the focus on novelty in NLP4RE research in which tools are used pre-
dominantly for publication purposes.

5.2.3 Motivation-Related Challenges.

Rec6. Tool reconstruction is not (yet) valued as a self-standing research contribution;
hence, researchers are discouraged from replicating tools over time. Similar to the annota-
tion task, reconstructing a tool is not sufficiently rewarding from a community standpoint. While
some venues accept replications in the form of scientific evaluation paper types that can involve
tool reconstruction, such papers could still be harder to publish. The reason for that, according
to the focus group's participants, is due to the RE community rating novelty aspects higher than
consolidation of knowledge.

6 DESIGN AND EVALUATION OF THE ID-CARD (RQ3)

Once the challenges related to data annotation and tool reconstruction were identified through
the responses to RQ1 and RQ2, we conducted Step ④ of our research method and implemented
the concept of ID-CARD.

6.1 ID-CARD Design

We designed the structure of the ID-CARD following an iterative method composed of three steps:
(1) Preliminary Definition, (2) Internal Assessment, and (3) External Assessment.
Preliminary Definition. In the first step, we outlined a list of information items needed for re-
constructing an NLP4RE solution based on our own experience in reconstructing the two tools
introduced in Section 4, as well as on the challenges gathered while answering RQ1 and RQ2. The
researchers worked in pairs, and each pair drafted a set of questions and possible answers for a
specific dimension. The considered dimensions included task, dataset, annotators & annotation
process, tool, and evaluation. The dimensions were selected based on previous experience and
brainstorming between the researchers. After the pairs of researchers drafted the questions, we
carried out a group meeting for 1 hour to consolidate them. At this stage, the researchers listed 56
questions. Another two iterations were carried out during the Internal Assessment (Step 2, which
is described below) to reach a stable set of 47 questions that are simplified to be more understand-
able by possible target readers.
Internal Assessment. We analyzed 46 papers from the NLP4RE literature (5 of which were co-
authored by at least one of the authors of this article) and extracted replication-relevant data
according to the information items from the first step. We also included in our analysisthe two
papers that we used in our replication scenarios (see Section 4).

Following the collection of the 46 selected papers, we used the preliminary ID-CARD (56 closed
questions); at least one researcher extracted information from each of the 46 papers. For four
papers, we also independently filled in the ID-CARD for the same paper to have some common
manual analysis in the same topic category, which we cross-checked and discussed at a later stage.
 The goal of this activity was to assess the applicability of the ID-CARD on a broad set of topics and
possibly customize it for different categories. After that, the researchers had a plenary discussion
to share the problems that emerged during this initial application of the ID-CARD and decide how

to mitigate these problems. The outcome of this meeting was an updated list of 47 questions in total, including a combination of 32 closed and 15 open questions.

External Assessment. The ID-CARD was shared with the authors of the original papers. We asked them to fill in the ID-CARD for their papers without sharing our previously filled-in cards. We contacted 32 authors, avoiding contact with the same research group for multiple papers, and received filled-in ID-CARDs from 15 authors for 15 different papers. For each paper, we analyzed the discrepancies between our input and the input of the original authors and further provided explicit notes listing the points of disagreement and reflections on possible motivations. In a 1.5-hour plenary session, we discussed the identified discrepancies based on the researchers' notes to identify major sources of disagreement. Most of the disagreements were due to the level of details provided. Different answers were provided with different levels of granularity to some questions in the ID-CARD. For example, in the case of multiple datasets used in the same study, some answers focused only on one dataset whereas others provided information about all datasets. Other disagreements were observed in the NLP task used in the study. In this case, we acknowledge that multiple NLP technologies can be applied in the same paper for solving some RE task. The goal of our plenary session was to identify the elements that could lead to misunderstanding, gather external viewpoints on the developed card, and discuss possible residual issues. After this analysis, we rephrased some of the questions according to our observations to make them more concrete and reduce possible disagreements.

6.2 ID-Card Description

The ID-CARD resulting from the various iterations consists of 47 questions, divided into 7 sections, described below. A compact version of the ID-CARD is displayed in Table 2. The card contains questions that either attempt to fully or partially address the challenges introduced in Section 5 or aim to provide metadata about the research paper, such as questions I.1 and II.1. In an online appendix [1], we provide the complete ID-CARD, validation material, and two filled-in ID-CARDs of the reconstructed tools described in Section 4.

I. RE Task. This section identifies the RE task addressed in the article, e.g., classification, tracing, and defect detection. The ID-CARD provides several options, from which only one option can be selected. The options include the nine categories listed earlier and an additional option that enables adding an unanticipated RE task. While the majority of NLP4RE papers address one main RE task, this is not always the case. For example, a paper could describe multiple RE tasks, such as generating a model from requirements specifications and a completeness checking method to identify incompleteness in requirements according to the generated model. Each RE task can be solved using a combination of NLP tasks such as text classification, named *entity recognition*, and semantic role labeling. Some papers also introduce different datasets and develop multiple tools. In this case, the ID-CARD is intended to be filled in separately for each of the RE tasks. The rationale behind this decision is two-fold. First, we simplify the overall design of the ID-CARD (e.g., we avoid the reiteration of certain questions for each RE task). Second, by decomposing the work into distinct RE tasks, we facilitate the retrieval of replication-relevant information associated with the paper, which is the main goal of the ID-CARD. Even when the entire work in such a paper is considered for replication, decomposing the work into distinct RE tasks helps better understand and replicate the tools more accurately.

II. NLP Task(s). This section specifies the NLP tasks used to support the RE task, i.e., classification, translation, information extraction and information retrieval (again, unanticipated NLP tasks may be specified). NLP tasks are distinguished from RE tasks as some NLP tasks could be applied for different purposes. For instance, classification (NLP task) can be used

Table 2. ID-card for NLP4RE Research Papers

	I. RE Task
I.1.	**What RE task is your study addressing?** ○ Requirements classification ○ Requirements tracing ○ Requirements defect detection ○ Model generation ○ Test generation ○ Requirements retrieval ○ Information extraction from legal documents ○ App review analysis ○ Dependency and relation extraction ○ Information extraction from requirements ○ Other . . .
	II. NLP Task(s)
II.1.	**What types of NLP task is your study tackling?** □ Classification (choose among classes) □ Translation (models, tests, etc.) □ Information extraction (glossary, terms) □ Information retrieval (search/rank) □ Other . . .
	III. NLP Task Details
III.1.	**What is the input of your NLP task?** □ Document □ Paragraphs □ Sentences □ Phrases □ Words □ Structured/tabular text □ Other . . .
III.2.	*[Only for Classification]* **What type of classification is the study about?** □ Binary-single label □ Binary-multilabel □ Multiclass-single label □ Multiclass-multilabel □ Other . . .
III.3.	*[Only for Classification]* **What are the labels that can be assigned?** . . .
III.4.	*[Only for Information Extraction]* **What is the level of granularity of the extracted elements?** □ Sentences □ Phrases □ Words □ Other . . .
III.5.	*[Only for Information Extraction]* **What is the type of the extracted elements?** . . .
III.6.	*[Only for Translation]* **What is the type of output?** □ Text □ Table □ Graphical diagram □ Executable model □ Test cases □ Other . . .
III.7.	*[Only for Translation]* **What is the translation mapping cardinality between initial input and final output?** □ 1 to 1 □ 1 to many □ Many to 1 □ Many to many □ Other . . .
III.8.	*[Only for Information Retrieval]* **What are the types of elements retrieved with the query?** □ Document □ Paragraphs □ Sentences □ Phrases □ Words
III.9.	*[Only for Other NLP Task]* **What is the type of output?** . . .
	IV. Data and Dataset
IV.1.	**How many data items do you process?** . . .
IV.2.	**In which year or interval of years was the data produced?** . . .
IV.3.	**What is the source of the data?** □ Industrial project, proprietary data □ Industrial project, publicly available data □ Community-based open-source projects □ Textbook examples or cases □ Student projects □ Toy requirements □ Legal/regulatory documents □ User-generated content □ Other . . .
IV.4.	**What is the level of abstraction of the data (not limited to requirements)?** □ User-level □ Normative-level □ Business-level □ System-level □ Module-level □ Code-level □ Other . . .
IV.5.	**What is the format of the data?** □ "Shall" requirements □ User stories □ Use cases □ User reviews □ Social media posts □ Bug/defect reports □ Messages in user forums □ Other . . .
IV.6.	**How rigorous is the format of the data?** □ Unconstrained natural language (NL) □ Template-based controlled NL □ Restricted grammar-based controlled NL □ Semantically augmented NL □ Other . . .
IV.7.	**What is the natural language of the data (if applicable)?** . . .
IV.8.	**Please list which domains your data belongs to:** *[comma separated]* . . .
IV.9.	**How many different sources does your data come from?** . . .
IV.10.	**Is the dataset publicly available (also from other authors)?** ○ Fully ○ Partially ○ Upon request ○ No ○ Other . . .
IV.11.	**What license has been used?** ○ No license ○ Reuse only for non-commercial purposes ○ Reuse for any purpose ○ Modification only for non-commercial purposes ○ Modification for any purpose ○ Other . . .
IV.12.	**Where is the dataset stored?** ○ On a private/corporate website ○ In a repository such as GitHub ○ In a persistent platform with DOI ○ Other . . .
IV.13.	**Provide a URL to the dataset, if available, or to the original paper that proposed the dataset** . . .
	V. Annotators and Annotation
V.1.	**How many annotators have been involved?** . . .
V.2.	**How are the entries annotated?** ○ Multiple annotators per entry ○ One annotator per entry (no quality control) ○ One annotator per entry (quality control) ○ Partly multiple annotators, partly one annotator per entry ○ Other . . .
V.3.	**What is the average level of application domain experience of the annotators?** ○ None or unknown ○ Informed Outsider ○ Domain expert ○ Other . . .
V.4.	**Who are the annotators?** ○ The designer of the technique/tool ○ People who have direct contact with designers ○ Independent annotators (e.g., crowd workers) ○ Other . . .
V.5.	**How was the annotation scheme established among the annotators?** ○ Only via labels ○ Oral agreement among the annotators ○ Written guidelines with label definitions ○ Written guidelines with definitions and examples ○ Other . . .
V.6.	**Did the authors make the written guidelines public?** ○ YES ○ NO
V.7.	**Did the authors share other information that could support the annotators other than the elements to annotate?** ○ NO ○ Surrounding context ○ Entire document ○ Other . . .
V.8.	**Did the authors employ techniques to mitigate fatigue effects during the annotation sessions?** ○ YES ○ NO
V.9.	**What are the metrics used to measure intercoder reliability?** ○ Cohen's Kappa ○ Krippendorf's Alpha ○ Fleiss' Kappa ○ Other . . .
V.10.	**How were conflicts resolved?** □ Majority voting □ Discussion among annotators □ Resolution by authors □ Resolution by independent expert (not an annotator) □ Disagreements were disregarded □ Not resolved □ Other . . .
V.11.	**What is the measured agreement?** . . .

(Continued)

Table 2. Continued

VI. Tool	
VI.1.	**What is the type of proposed solution?** □ Supervised machine learning (ML) □ Unsupervised ML □ Supervised deep learning (DL) □ Unsupervised DL □ Rule-based □ Other . . .
VI.2.	**What algorithms are used in the tool?** . . .
VI.3.	**What has been released?** □ Standalone tool □ Service on the web □ Library/API □ Source code □ Executable notebook □ Pre-trained models □ No tool has been released □ Other . . .
VI.4.	**What needs to be done for running the tool?** ○ No installation is needed ○ Import and integrate into your own code ○ Compile and run ○ Virtual machine/Docker container ○ Reproduce the tool from the explanation of the paper ○ Other . . .
VI.5.	**What type of documentation has been provided alongside the tool?** □ README file □ Pseudocode/illustration in the paper □ Wiki or dedicated website □ Tutorial □ Ready-to-use examples □ An academic paper □ No documentation □ Other . . .
VI.6.	**What type of dependencies does the tool have?** □ None □ Open-source libraries/software □ Proprietary libraries/software □ Specific operating system □ Specific hardware □ External knowledge bases □ Other . . .
VI.7.	**How is the tool released?** ○ On a private/corporate website ○ In a repository such as GitHub ○ In a persistent platform with DOI ○ Other . . .
VI.8.	**What license has been used? Other** ○ No license ○ Reuse only for non-commercial purposes ○ Reuse for any purpose ○ Modification only for non-commercial purposes ○ Modification for any purpose ○ Other . . .
VI.9.	**Where is the tool released?** *[url]* . . .
VII. Evaluation	
VII.1.	**What metrics are used to evaluate the approach(es)?** □ Precision/recall □ F-score □ Accuracy □ AUC □ MAP □ LAG □ Word Error Rate (WER) □ NIST – METEOR – ROUGE – BLEU □ Other . . .
VII.2.	**What is the validation procedure?** ○ Train–test split ○ Cross-validation ○ Cross-project validation ○ Entire dataset ○ Other . . .
VII.3.	**What baseline do you compare against?** □ Existing tool or algorithm □ Reconstructed tool from other research papers □ Automated, but self-defined □ Theoretical/conceptual □ Human baseline □ None □ Other . . .
VII.4.	**Please provide more details about the baseline you compare against, if any.** . . .

to organize requirements into different categories but also to detect defects or identify trace links (RE tasks). More than one NLP task can be selected since the NLP tasks can be used in combination—e.g., in model generation, one uses information extraction *plus* translation.

III. NLP Task Details. This section characterizes the details of the RE task addressed, i.e., the RE task input granularity (e.g., document, word, paragraph) and the output type. We define various options of the output types that differ for each NLP task. For example, a classification task requires specifying whether the output is binary-single label (e.g., ambiguous XOR not ambiguous), multi-class multi-label (e.g., feature request OR bug OR praise), or other options in between. One has to further specify the possible labels (i.e., classes) of the output. For translation, the output can be text but also test cases or models. Specific to translation, one has to specify the *cardinality*, i.e., one input to one output (e.g., one document to one diagram) many inputs to many outputs (e.g., from many use case descriptions to one class diagram and multiple other diagrams).

IV. Data and Dataset. This section characterizes the dataset used in the article. One has to provide details about the size of the dataset; the year; the raw data source data (e.g., proprietary industrial data, regulatory documents, user-generated content—eight options plus "other"); the level of abstraction of the data (e.g., user-level, business-level, system-level). In this section, the term *data* collectively identifies both possible input and output data of the previously selected NLP task(s). Therefore, the questions support multiple answers. This choice was driven by the need to keep the ID-CARD well structured and easy to retrieve. This section also asks for information concerning the format of the data (use case, "shall" requirements, diagrams, etc.), the degree of rigor (unconstrained NL, restricted grammar, etc.), and the actual language (if applicable). Additional questions are included about the heterogeneity of the dataset—in terms of domain coverage and number of sources from where the data is obtained—as well as about the data licensing and the URL to access the dataset.

V. Annotators and Annotation Process. This section includes information to characterize the annotators in terms of background knowledge, number, and level of bias in case annotation was carried out on the raw data. In addition, information is collected about the adopted annotation scheme (if any) and the process to measure and resolve disagreements. This section focuses mainly on manual annotation, a common practice for creating datasets in NLP4RE. With the current NLP technologies, researchers are shifting towards creating datasets using automated means. Extending the annotation section to cover the replication of automated annotation is left for future work.

VI. Tool. This section collects information about any implementation provided along with the article, e.g., scripts, executable programs, **application programming interfaces (APIs),** collectively designated with the term *tool*. Specifically, the questions in this section require information about the enabling technology of the implemented NLP solution (e.g., machine learning, rule-based), what has been released (e.g., binary file, source code), and additional information concerning documentation, licensing, dependencies, and other details that can help access and execute the tool.

VII. Evaluation. This section requires information about the evaluation carried out in the article: the evaluation metrics (precision/recall, **Area under the Curve (AUC)**, etc.), the type of validation process (cross-validation, train-test split, etc.), and the baselines used for comparison (if applicable). Investigating other evaluation alternatives, e.g., the impact of using the tool on the downstream development, is left for future work.

We note that the ID-card provides a comprehensive view of what information is relevant for replication. However, for a particular paper, one might fill in only some sections of the ID-card that are found in and relevant to the paper. For example, if the paper only presents a new dataset without an automated approach, then only the section about annotation might be relevant.

6.3 Tracing Challenges to Questions in the ID-card

In the following, we discuss the rationale for the majority of the questions in the ID-card, in relation to the challenges described in Table 1. We note that the traceability is not one-to-one mapping between challenges and the ID-card. Some challenges, such as the little value given by our RE community to replicating studies, cannot be handled by the ID-card.

Recall that the classification of the RE tasks (Question I.1) was adopted from the recent systematic literature review by Zhao et al. [57], whereas the NLP task classification (Question II.1) was specifically designed to complement the RE tasks. The ID-card partially addresses the annotation challenges (see Table 1) as follows:

— To address challenge Ann1, questions V.5 and V.6 are introduced to highlight the necessity for establishing a clear annotation protocol prior to the annotation process and further making it publicly available afterwards. Question V.10 also points out that communicating among annotators is required for achieving a high-quality dataset. Once the annotation protocol is agreed on, the likelihood that it frequently evolves (Ann4) is limited.

— While including domain experts in the annotation process is not always possible, questions V.1 to V.4 in the ID-card inform the researchers interested in reconstructing a dataset about whether the domain expertise was involved or not.

— With regard to Ann3, Question V.8 indicates the need to mitigate fatigue.

— Questions IV.10 to IV.13 are concerned with the dataset. The questions spotlight the publicly available datasets along with the licenses under which they are released. By leveraging such datasets in future similar annotation tasks, one can address the challenges Ann5 to Ann7.

(a) ID-CARD perceived ease of use.

(b) ID-CARD intention to use for different use cases.
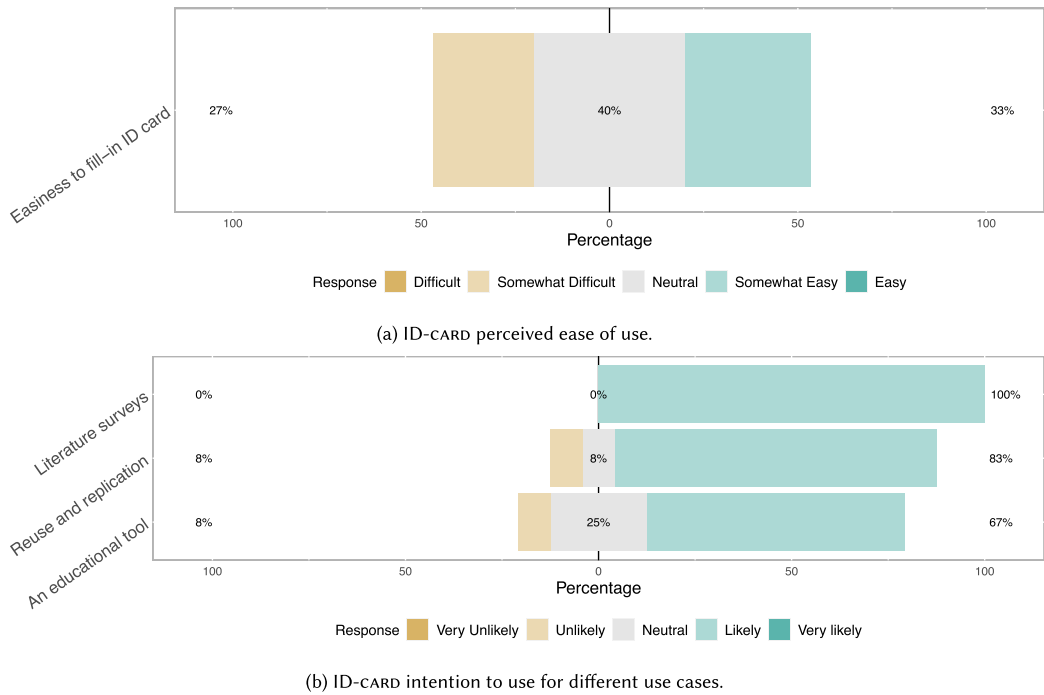
Fig. 3. Results of TAM assessment for the NLP4RE ID-CARD.

— Question V.7 gives insights about the context that is shared with the annotators during the annotation process. The ID-CARD can be used to create a common practice to assist researchers in designing new annotation tasks, thus, addressing the challenge Ann8.

With regard to the tool-reconstruction challenges, the ID-CARD demands more details and precise information concerning the implementation of the tool which are often omitted from a paper due to space limitations. Specifically, Questions VI.6 and VI.8 are about the libraries employed in the implementation, Question VI.5 is about the available documentation related to the tool, and Questions VI.3 and VI.9 are about what has been publicly released.

6.4 Evaluation of the ID Card

We disseminated a short survey (four Likert-scale questions and two open-ended questions) among the 15 authors who filled in the ID-CARD, asking them to provide explicit feedback. Inspired by the questionnaire from the **Technology Acceptance Model (TAM)** [17], we included survey items (see Figure 3) regarding the perceived ease of use and intended use in three different use cases, reuse and replication, literature surveys (as typical instrument that NLP4RE newcomers use for learning the state-of-the-art), and education. The vast majority of respondents agreed that the ID-CARD has the potential to be used for the purposes stated above. Nevertheless, opinions were mixed regarding the card ease of use. In the open feedback section of the survey, some respondents indicated a preference for the support of multiple datasets and a more interactive format that support conditional sections based on previous answers.

We also evaluated the respondents' perception about the ID-CARD appropriateness and level of details for the three different use cases (see Figure 4). Although the vast majority agrees that the use of the ID-CARD is appropriate across the three use cases, they perceived that more details may

(a) NLP4RE ID Card perceived appropriateness for different use cases.

(b) NLP4RE ID Card perceived level of details.
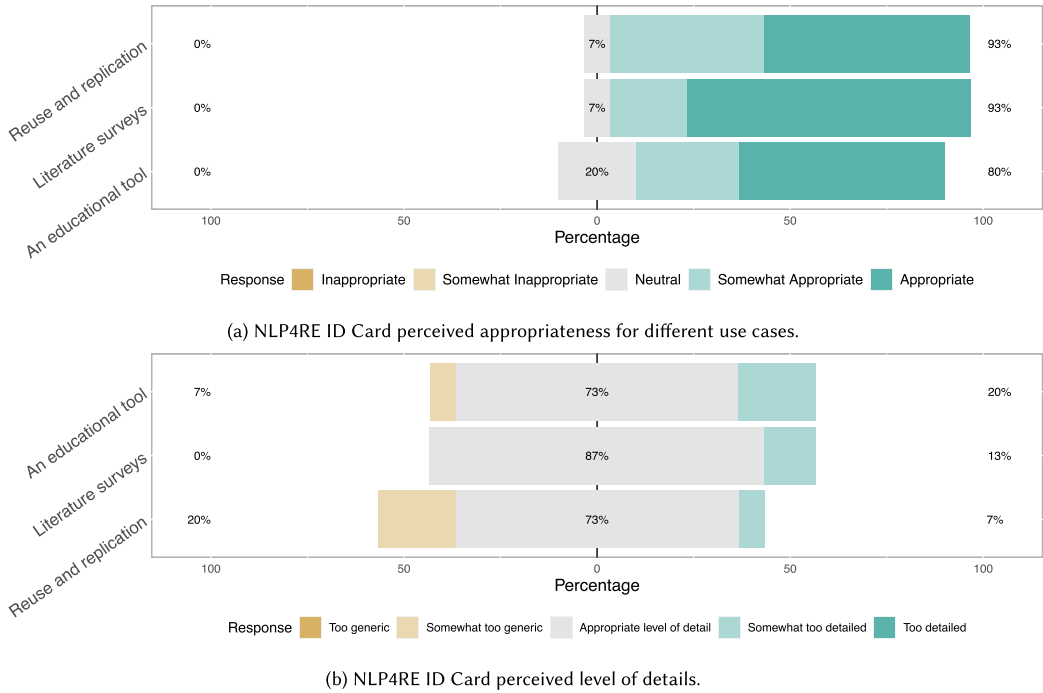
Fig. 4. Assessment of the NLP4RE ID Card appropriateness and level of detail.

be necessary to further facilitate reuse and replication of existing work. None of the respondents indicated which details should be included, but the necessity to strike a balance between details and usability was recognized, as one of the respondents pointed out: *"For replication purposes, I think more details would be required. On the other hand, putting in more detail will make the ID-CARD creation more time-consuming (potentially defeating the purpose). There is no easy answer here."*

The respondents indicated other possible usages, such as leveraging the information contained in the card for searching and filtering previous literature, e.g., when selecting a baseline to compare in their own research work, and as a checklist to follow during the planning phase of a new work.

7 DISCUSSION

In the following, we discuss how the ID-CARD mitigates the identified challenges, relevant hints to fill the ID-CARD, lessons learned, and limitations of the study based on the discussion carried out in the final focus group (see Step ⑥ in Figure 2).

7.1 Mitigation of Challenges

In Section 5, we have identified several challenges concerning tool reconstruction and dataset annotation. To mitigate the above-listed challenges or reduce their effect, we suggest below some mitigation actions.

— **Rigorous Annotation.** Authors should make the annotation task more rigorous by (i) conducting pilot annotations to set the protocol and to avoid later changes; (ii) maintaining written guidelines, including examples, to guide the annotators as well as for future reference; (iii) performing reconciliation sessions to resolve disagreements; (iv) reminding the annotators about their ethical responsibilities—the annotation should be as reliable as possible, as it may be reused by other researchers in other work.

— **Reward reconstruction.** The research community should consider rewarding annotation and tool reconstruction tasks by creating more venues and events where such activities can be published and shared with the community.

— **Clarity.** The natural language description of a tool present in a paper is often insufficient and too informal for other researchers to reconstruct the tool unambiguously. Therefore, researchers should allocate sufficient effort to clearly describe how the tool can be reconstructed from scratch.

— **Flexibility.** Flexibility is essential and often necessary to properly reconstruct existing tools. This includes simplifying or adapting existing tools to be applicable in the reconstruction scenario, e.g., using different data, filling in missing implementation details, replacing proprietary or legacy libraries.

— **Goal-driven Reconstruction.** The rigor of reconstruction depends on the *goal* of the endeavor. For instance, building a baseline to compare a new solution requires more rigor than for reconstructing a tool for practical use in a company.

7.2 Guidelines for Filling the ID-CARD

One of the outcomes of the third focus group the authors held was the identification of a couple of considerations related to filling in the ID-CARD. In particular:

— **One ID per task.** The first step when filling in the ID-CARD is to map the proposed NLP solution to the underlying RE task. However, many papers discuss multiple NLP solutions (e.g., extraction and classification) in one pipeline for solving the same RE task (e.g., solving cross-referencing in requirements). To improve re-usability and foster reconstruction, we recommend filling in the ID-CARD by focusing on one RE sub-task, which is likely the main task being solved in the paper. The ID-CARD in its current status can be theoretically filled in multiple times, each time for an RE sub-task.

— **Degree of detail.** To identify the reconstruction-relevant information, it might not be clear for the researcher filling in the ID-CARD to what extent such information should be sought beyond the original paper. Although having all possible details is the best option for reconstruction, this poses a pragmatic challenge related to the additional time required for the respondents to fill in the ID-CARD. The researcher should address this consideration in line with the motivation of filling in the card. For example, the author of a paper could provide more details for better reuse and replication (see Figure 3(a)). We acknowledge that the ID-CARDs are not expected to be identical for the same paper when filled in by two different researchers, yet they would still represent equivalent summaries of the original paper. However, filling in multiple ID-CARDs for the same paper is unlikely in future practice when the authors provide an ID-CARD along with their publications.

7.3 Lessons Learned on the ID-CARD Design

Another outcome of the final focus group was several lessons learned during the design of the ID-CARD, which can be useful for researchers addressing similar endeavors. Among them, we offer the following remarks.

— **Generality of the ID-CARD.** The details required by the card should cover sufficient information about the solution without merely repeating what is in the paper. The ID-CARD aims to summarize an NLP4RE paper to replicate the solution and/or dataset. Ensuring that it can cover a wide spectrum of the papers in the NLP4RE literature is important, yet challenging. There is a trade-off between the coverage of the paper by the ID-CARD versus the level of detail required to achieve a comprehensive description of the paper. Requesting many

details entails more time and effort filling in the card. To address this consideration, we opted to design the card at a generic enough level to be applicable to many papers of different research focus. The ID-CARD should help researchers decide whether or not papers are useful for their reconstruction goals.

— **Free text options.** Missing details that are relevant to reconstruction must be properly accounted for and incorporated in the ID-CARD. We address this consideration by adding a free text option alongside each question to give the possibility to elaborate on the reason or remarks concerning missing details. We believe that providing a justification can give hints about what to do with these missing details. For example, if the evaluation uses the entire dataset instead of cross validation, the reason might be that the tool does not require training. In this case, details about the proportion of training data might be missing but also not needed.

7.4 Possible Uses of the ID-CARD

The ID-CARD serves several purposes to different actors, as summarized in Table 3. To NLP4RE newcomers, e.g., researchers entering the field and PhD students starting their thesis, the ID-CARD is an effective instrument to get acquainted with the state-of-the-art. Experienced researchers may use it when conducting their NLP4RE research at several stages—namely, design, evaluation, and reporting—as a checklist to ensure that their study covers all relevant aspects required to be replicable. In the frequent case that the work is presented in a paper, researchers may submit the ID-CARD to a public repository as accompanying material. This way, reviewers of the paper have access to more detailed information, which helps provide an informed evaluation of the paper. Last, we foresee the ID-CARD as a useful source of information for educators when preparing NLP4RE-related material.

We further envision that the ID-CARD can be used to create and archive summaries of papers which one has to review at a particular time and come back to them again later—e.g., researchers performing snowballing for a literature review can quickly look for the search seeds among such an archive.

Finally, the ID-CARD can be used to assess the adequacy for the reconstruction of a particular solution without having to read the entire paper, e.g., incomplete description, missing implementation details. In fact, it increases the chances of reconstructing a paper. In the future, ID-CARDs can be collected from the authors during the paper submission/review process and stored in a public repository to be publicly accessible by the RE community. Authors and reviewers can also use it as a checklist of all the information that should be reported in the paper concerning reconstruction and reusability. Thus, the ID-CARD can create more awareness of including such necessary information.

7.5 Limitations

This section identifies the limitations of our study and discusses how we mitigate them. We group the limitations, distinguishing those related to the identification of the challenges and to the construction of the ID-CARD.

The problems of requirements ambiguity detection and requirements classification are classical and widely studied topics in NLP4RE, as shown by the survey of Zhao et al. [57]. However, the specific cases discussed during the focus groups are selected based on the experiences of the authors; thus, personal *bias* might threaten the validity of the results. To mitigate personal bias, experts in different NLP4RE tasks participated in the focus groups and shared their views. These focus groups involved expert NLP4RE researchers who were involved in the two presented replications; moreover, all the experts had extensive experience in various NLP4RE tasks. The participants also

Table 3. Actors Using the ID-CARD as Support for a Number of Activities

Actor	Activity	Use of the ID-CARD
NLP4RE newcomer	Learn state-of-the-art	Having a quick introduction to NLP4RE tasks using the ID-CARD structure
		Knowing the characteristics of existing solutions using the ID-CARD contents
Study author	Design the study	Learning the relevant aspects to be covered in the study using as checklist the ID-CARD structure for the target type of NLP4RE task
	Evaluate the study	Searching and filtering state-of-the-art tools that act as baseline
		Reusing or reconstructing those tools
		Comparing results with respect to baseline in a systematic way
	Report the study	Summarizing the detailed characteristics and results of the study
Paper author	Submit paper	Submitting the ID-CARD as accompanying artifact for the paper
Paper reviewer	Read paper	Using the ID-CARD contents to complement the paper contents
	Evaluate paper	Using the ID-CARD structure as a checklist for the evaluation
	Write review	Using the ID-CARD structure to structure the review
Educator	Prepare course materials	Organizing course materials using the ID-CARD structure
		Summarizing the state-of-the-art using the ID-CARD contents

identified commonalities among the challenges identified in different focus group studies, which indicates that the challenges were not identified based on the personal bias of the participants. In order to prevent any threats due to *social relations and persuasion*, each focus group was moderated by a different person, and had distinct discussion leaders. During the discussion, each participant was given the opportunity, and encouraged, to freely speak up. Nevertheless, some participants were more vocal than others.

During the focus groups, identified challenges may not be mapped to the same category in the cognitive state of the participants. Thus, *understandability* of the challenges may be a threat. To mitigate this, each moderator took notes during the focus group and shared those notes in real time so that the other participants could read and clarify them when necessary. Each moderator summarized the findings at the end of each question during the focused discussion phase and also at the end of the focus group. *Completeness* of the challenges identified in the focus groups is another threat. To help achieve completeness, we did not enforce time limitations on the focus groups, allowing the participants to deeply discuss the cases and to add new points as they saw fit. To ensure *conciseness*, we iterated over the identified challenges to check for redundancy and reduced the redundancy. The challenges are identified based on two distinct, yet representative, NLP4RE cases. It is worth remarking that the two cases were used as triggers for the discussion, and the actual challenges elicited are based on the analysis of 46 papers and on the substantial expertise of the participants of the focus groups, who have participated in several artifacts tracks, in NLP4RE tool design, and in study replications. Further experimentation is however needed to ensure *generality* of our findings. Some challenges may apply to the field of **natural language processing for software engineering (NLP4SE)**, whereas others may be more general to SE. Nevertheless, as the scope of this study is strictly NLP4RE, more studies are needed to support these claims.

Two main concerns for the ID-CARD are *completeness* and *pragmatic usability*, which may conflict with each other. We tried to find a healthy balance between the two using multiple iterations of application and refinement of the ID-CARD, although we expect that lightweight versions of the card may be created for specific tasks and uses, e.g., for those listed in Table 3. These also increased *understandability* of the card as we applied it to multiple articles. After each iteration, we discussed the results to confirm that a shared understanding was reached. Understandability was further assessed by asking the authors of the paper to fill in the ID-CARD. To prevent the personal *bias* of

single authors affecting the construction of the ID-CARD, multiple experts of the different NLP4RE tasks participated in its construction. Issues related to *coverage* of the ID-CARD in terms of types of alternative answers for each question was mitigated by assessing the card on 46 papers specifically selected to cover the spectrum of typical RE tasks based on the taxonomy by Zhao et al. [57] and considering both widely cited seminal papers and relevant recent ones.

8   CONCLUSIONS AND OUTLOOK

Replication, covering both data (re-)annotation and tool reconstruction, is an important strategy in experimentation and empirical evaluation. In this article, addressing the field of NLP4RE, we investigated what are the challenges of annotating datasets for training and evaluating NLP4RE tools (RQ1) and the challenges in reconstructing NLP4RE tools (RQ2). To answer these research questions, we conducted focus groups in which we reflected on our first-hand experience in replicating NLP4RE state-of-the-art tools, and we further analyzed 46 papers covering a wide spectrum of the NLP4RE landscape. As a result of our study, we identified 10 challenges concerning data annotation and 5 challenges for reconstructing tools. Some of these challenges are specific to NLP4RE whereas some can be considered applicable to other SE fields. Though we refrain from generalizing our findings, as they stem from an NLP4RE context, we encourage other authors to further investigate our list of challenges and possibly adapt it to their contexts.

Challenges concerning data annotation (RQ1) include the unavailability of theories specific to RE tasks to build the annotation task on, e.g., concrete definitions of non-functional requirements or nocuous ambiguity; the need to anticipate additional time and effort to potentially evolve annotation protocols; and issues resulting from dataset imbalance and the lack of domain knowledge of annotators. Challenges for reconstructing tools (RQ2) mostly arise from missing details in the original papers, e.g., the exact library version or the application of outdated NLP tools that are no longer available.

To reduce the effect of these challenges, we investigated how to support NLP4RE researchers (RQ3). We proposed an ID-CARD as a complementary source to original papers for summarizing, via a total of 47 questions, information relevant to replication. The ID-CARD covers seven topics essential for replication. These concern the RE and NLP tasks, inputs, outputs, annotation process, tool, and evaluation details. We assessed the ID-CARD both internally and externally (with the authors of original papers) by cross-filling the same paper. Though filling in the ID-CARD requires time and effort (which should be marginal for the authors of a paper), the ID-CARD provides a useful starting point to facilitate replication.

In the future, we would like to foster the ID-CARD as part of paper submission at different RE venues. The goal is to help reviewers assess the replication consideration of the paper and increase the overall awareness of replication-relevant information. Furthermore, we plan to investigate extending the ID-CARD to cover other SE-related areas. This is highly needed, as artifact evaluation is not sufficiently mature and needs further improvement [32, 53]. Though the ID-CARD mainly aims to improve artifact evaluation, this is not its only purpose. It can be used for other objectives, such as literature reviews, as also indicated by the participants in our external assessment. These are hypothetical scenarios of usage, which can evolve after the ID-CARD is used and possibly adapted by the community.

APPENDIX

A   DETAILS OF THE FOCUS GROUPS

Protocol. Table 4 reports the moderators, lead participants, and regular participants for the three focus groups. All authors are experienced in the annotation and tool reconstruction tasks. We follow the guidelines by Breen [6]:

Table 4. Moderators, Leads, and Participants of the Focus Groups

	Moderator	Leads	Participants
Focus Group RQ1	A2	A1, A3, A4, A6	A5, A7
Focus Group RQ2	A5	A1, A2, A4	A3, A6, A7
Focus Group ID-CARD	A6	A1, A3	A2, A4, A5, A7

Focus Group Preparation. This task includes the following steps.
 (1) A moderator is appointed among the authors who did not participate in the annotation (Focus Group 1) or reconstruction cases (Focus Group 2), or ID-CARD definition (Focus Group 3).
 (2) Lead participants are selected among those who participated in the annotation, reconstruction, and ID-CARD definition cases. Regular participants are those who are not lead participants or moderators.
 (3) A meeting is organized in which the moderator is informed about the results by the lead participants and is given an overview of the collected data.
 (4) The moderator defines a schedule for the focus group and organizes it.
 (5) Before each focus group, the lead participants are required to independently prepare a list of challenges or relevant discussion topics about the experience. These are shared during the focus group execution.

Focus Group Execution. The focus group session is planned for 90 minutes: 40 minutes for free discussion, 40 minutes for focused discussion, and 10 minutes for wrapping up the session. The moderator starts the session by introducing the main question to be addressed by the focus group, i.e., the RQ associated to the focus group. Following this, the lead participants present their experience to the others in two slides prepared in advance prior to the actual session. The moderator then collects a list of challenges mentioned by the lead participants in one slide and moderates the discussion among the participants in a free-form manner. Further questions are progressively added and addressed within 8 minutes each during the focused discussion. After the focused discussion, the moderator summarizes the discussion and concludes the session.

Qualitative Data Collection and Analysis. The moderator of the focus group analyzes the tape recording and extracts the themes that were discussed the most and perceived as the most important. This is done by the moderator to ensure neutrality, as the moderator is someone who did not directly participate in the activity discussed in each focus group.

Review of Qualitative Data. The themes are proposed to the participants in a wrap-up meeting. This is conducted 2 weeks after each focus group, lasts 9 minutes, and involves all participants. The participants can complete, enrich, or amend the themes to come to an agreed set of challenges around the specific RQ.

Dataset Annotation and Re-annotation Focus Group Details This first focus group aimed to discuss the challenges related to data annotation for NLP4RE tasks as described in Section 3. It started with the presentations of the lead participants for the free discussion. During the free discussion, 18 challenges were listed by the participants and documented by the moderator in a shared document. The free discussion was followed by the questions asked by the moderator for the focused discussion. Below, we list these questions and summarize the discussion items.
 — **Q:** How do you decide on the labels for the data? **A:** Several factors affect the decision on the labels: (i) existing theory or (ii) the limitations of the state-of-the-art can be the sources for the set of labels. At the same time, (iii) the recency of the labels used in the state-of-the-art is a determining factor on the decision whether to update the set of labels.

Another factor is (iv) the availability of the raw data: following an opportunistic approach, the researchers may add or drop some labels. Finally, (v) assumptions from linguistic and industrial backgrounds may have an impact on the set of labels (e.g., linguistics ambiguity does not necessarily lead to multiple interpretations in software development due to company conventions and standards).

— **Q:** How do you construct the annotation protocol? **A:** Protocol construction is an iterative process. An initial (pilot) annotation helps to construct a protocol. The annotators need to seek a balance between the effort put in the annotation process and the cost of validation and re-annotation resulting from potential changes in the protocol. The participants also observed that the protocols do not guarantee an agreement among the annotators, and even when the protocols are documented as written guidelines, human aspects such as fatigue handling or persuasion avoidance are typically not documented in the annotation protocol. This answer summarizes the experience of the authors on annotating data for NLP4RE tasks.

— **Q:** How do you validate the labels and annotation protocol? **A:** The validation is done by checking the labels, and the protocols are aligned with the purpose of the annotation process. Due to practical considerations, researchers typically proceed with a good-enough protocol. The consensus is that the annotators do not have strict protocols as in medical studies. Annotation protocols are more flexible and prone to change during the annotation process. Since trials and changes are costly, the researchers are content with good-enough annotation protocols and do not strive for the perfect ones.

— **Q:** What are the requirements for the annotators? **A:** Independence of the annotators is a desired quality, which is difficult to satisfy unless there is a budget for the annotators. For some annotation tasks, domain knowledge can be crucial to ensure a reliable outcome. Experience in annotation, whether for RE or not, is also desired. When seeking skilled annotators, crowd-sourcing platforms may output annotations with questionable quality. Doctoral students, on the other hand, can efficiently provide good-quality results.

— **Q:** How are the annotators trained? **A:** Written guidelines are used as the training material. Pilot sessions and annotating in small iterations also contribute to the training process. Yet there is no formal training on the group theory [33] and collaborative decision-making. Those skills are acquired through experience.

— **Q:** What are the format and tools used for annotation? **A:** Practical concerns determine the tool for the annotation process and the format of the annotated data. Simple tools such as spreadsheets and common file formats such as **comma-separated values (CSVs)** are preferred.

— **Q:** When do you stop annotation? **A:** The size of the annotated data, class balance—having a comparable number of samples for each class—and practical concerns such as the availability of the annotators can be important factors for this decision, which is taken during the process.

— **Q:** Is annotation a rewarding experience? **A:** The focus group participants found the annotation process both rewarding and unrewarding. While the annotation process itself can be interesting and triggering for some, providing a deeper understanding of the problem and the research as well, it may quickly become tedious and not challenging, especially with the increase of the size of the dataset. Nevertheless, the participants found that sharing the dataset could be rewarding, as it is used by the community.

Tool Reconstruction Focus Group Details. This focus group aimed to discuss and reflect on our first-hand experience in reconstructing the two state-of-the-art solutions introduced in Section 4.

The moderator started the session by introducing the main question to be addressed by this focus group, which is "What are the challenges in reconstructing NLP4RE tools?". Next, the lead participants presented their experience to the others in two slides prepared in advance prior to the actual session. The moderator then collected a list of 23 challenges mentioned by the lead participants in one slide and moderated the discussion among the participants in a free-form manner. The study continued with the focused discussion in which further questions were progressively added and addressed within 8 minutes each during the focused discussion. Below, we list the questions and summarize the discussion items.

- **Q:** Where do you start when reconstructing tools? **A:** For reconstructing a tool, one starts by (1) defining the goals (e.g., benchmark or baseline), (2) selecting a state-of-the-art paper that is clear enough, (3) identifying the tool components to be reconstructed, and (4) adapting the tool to fit the available dataset.
- **Q:** Do you follow a structured and repeatable process for tool reconstruction? **A:** Reconstruction is often an *ad-hoc* process depending on the underlying case, reconstruction goals, and what is publicly available (e.g., source code).
- **Q:** How do you ensure that the reconstruction is correct? **A:** Unless the original tool is publicly available, comparison is not always possible due to lack of implementation details (e.g., hyper-parameters or library versions). However, one should use the same evaluation metrics and validate intermediate steps using simple examples.
- **Q:** What are the most time-consuming phases of tool reconstruction? **A:** Adapting the original tool to fit the reconstruction goals requires deriving the implementation details from the original paper. Most often, time and effort are put into checking alternatives and making decisions about the unknowns (e.g., which parameter value, imbalance handling, thresholds).
- **Q:** Do you consider tool reconstruction a rewarding experience? (Would you do it again?) **A:** The participants agreed that reconstruction is challenging. At the individual level, it could be to some extent rewarding, as one gets the opportunity to learn about the process and spot incorrectness or conceptual errors. At a community level, reconstruction can be rewarding if the reconstructed tool is shared, e.g., as a baseline.

Finally, additional issues emerged during the focus group. In summary, dealing with the unknowns can be an important factor for determining the divergence level between the reconstructed tool and the original one. Interaction with the authors is another factor that can facilitate the reconstruction process; however, since maintaining a tool is not common within the academic research community, authors' help might be limited.

ID Card Design Focus Group Details The focus group session followed a similar structure to the previous ones. The seven researchers (co-authors of this article) participated in this focus group as one moderator, two lead participants, and four discussants.

In this case, the focus group was moderated by one of the authors (A6) who did not participate in the development of the ID-CARD. The moderator started the session by introducing the main question to be addressed by this focus group, which is "What are the challenges in extracting replication-relevant information from NLP4RE papers?". Following this, the lead participants presented their experience to the others in two slides prepared in advance prior the actual session. The moderator then collected a list of 21 challenges mentioned by the lead participants in one slide and moderated the discussion among the participants in a free-form manner. The challenges were grouped into five categories, defining content, designing questions and answers, evaluation and inter-rater agreement, paper-structure related and other. Further questions were progressively

added and addressed within 8 minutes each during the focused discussion. Below, we list the questions and summarize the discussion items.

— **Q:** Is there any aspect that the ID-CARD does not reflect? **A:** Some aspects were intentionally left out, e.g., some RE tasks or details concerning manual steps that involve interaction with users. This choice was motivated by three conditions. First, the ID-CARD should be generic enough to cover a wide spectrum of RE literature. Thus, instead of specifying all details, we substituted the omitted choices with open questions that allow free text answers. Second, the ID-CARD should not be regarded as a replacement of a paper; rather, it should be seen as a complementary resource that provides a structured summary. For example, context that is part of the paper should not necessarily be part of the ID-CARD. Finally, limiting the ID-CARD with a finite list of answers (choices) reduces its usability in other contexts, e.g., SE tasks that are not part of the NLP4RE literature.

— **Q:** Is it necessary to complement the provided answers with explanations, justifications, or any evidence? **A:** Though explanations are not enforced in the ID-CARD, they are still encouraged. We have added a free-text choice to each question to provide the tagger with the possibility to complement a given answer with further details, justifications, or evidence.

— **Q:** How trustworthy is the collected information? How do annotators' disagreements affect the overall quality of the evaluation? **A:** We observed that most disagreements are about the level of details that the annotators provided rather than actual contradictions. Thus, we believe that the ID-CARD is trustworthy as a companion to the paper that facilitates the access to replication information, the original purpose for which it is designed.

— **Q:** How can the ID-CARD be used in the context of replication? **A:** Since the ID-CARD provides a structured summary of the original paper, it informs researchers about the suitability of the study for their replication scenario. To increase the benefit of the ID-CARD, we envisage appending it as supplementary material to future papers and/or collecting the ID-CARDs in one public repository to be shared with the community. The ID-CARD can further provide guidelines for reviewing papers. Generalizing the ID from NLP4RE to other (more general) domains, e.g., NLP4SE, might require effort but is rewarding in the long term.

— **Q:** Do you consider filling in the ID-CARD a rewarding experience? (Would you do it again?) **A:** According to the participants in the focus group, filling in the ID-CARD would require less effort and time from the author of the paper.[10] Thus, the participants agree on the advantage of encouraging the creation of an ID-CARD to be publicly released along with the paper.

REFERENCES

[1] Sallam Abualhaija, Fatma Başak Aydemir, Fabiano Dalpiaz, Davide Dell'Anna, Alessio Ferrari, Xavier Franch, and Davide Fucci. 2024. Online Annex: Replication in Requirements Engineering: The NLP for RE Case. https://doi.org/10.6084/m9.figshare.21824481

[2] Sharif Ahmed, Arif Ahmed, and Nasir U. Eisty. 2022. Automatic transformation of natural to unified modeling language: A systematic review. In *2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA '22)*. 112–119. https://doi.org/10.1109/SERA54885.2022.9806783

[3] Carlos E. Anchundia and Efraín R. Fonseca C. 2020. Resources for reproducibility of experiments in empirical software engineering: Topics derived from a secondary study. *IEEE Access* 8 (2020), 8992–9004. https://doi.org/10.1109/ACCESS.2020.2964587

[4] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (2016).

---

[10]During the focus group, the participants have agreed on the hypothesis that filling in the ID card for a paper would require less time and effort if done by the author of that paper. Considering the responses on the ease of use of the ID-CARD (see Figure 3 in Section 5 of our article), we observed that this hypothesis is not always true, in particular when the author fills in the ID card for a paper that has long been published.

[5] D. Berry, E. Kamsties, and M. Krieger. 2003. From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity, A Handbook. Retrieved from http://se.uwaterloo.ca/~dberry/handbook/ambiguityHandbook.pdf

[6] Rosanna L. Breen. 2006. A practical guide to focus-group research. *Journal of Geography in Higher Education* 30, 3 (2006), 463–475.

[7] Jeffrey C. Carver. 2010. Towards reporting guidelines for experimental replications: A proposal. In *1st International Workshop on Replication in Empirical Software Engineering*, Vol. 1. 1–4.

[8] Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 18th Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*.

[9] Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, and Peter Solc. 2006. The detection and classification of non-functional requirements with application to early aspects. In *14th IEEE International Requirements Engineering Conference*. 39–48.

[10] Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, and Peter Solc. 2007. Automated classification of non-functional requirements. *Requirements Engineering* 12, 2 (2007), 103–120.

[11] Margarita Cruz, Beatriz Bernárdez, Amador Durán, Jose A. Galindo, and Antonio Ruiz-Cortés. 2019. Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018. *IEEE Access* 8 (2019), 26773–26791.

[12] Fabio Q. B. Da Silva, Marcos Suassuna, A. César C. França, Alicia M. Grubb, Tatiana B. Gouveia, Cleviton V. F. Monteiro, and Igor Ebrahim dos Santos. 2014. Replication of empirical studies in software engineering research: A systematic mapping study. *Empirical Software Engineering* 19, 3 (2014), 501–557.

[13] Fabiano Dalpiaz, Davide Dell'Anna, Fatma Basak Aydemir, and Sercan Çevikol. 2019. Requirements classification with interpretable machine learning and dependency parsing. In *Proceedings of the 27th IEEE International Requirements Engineering Conference, RE 2019*. 142–152. https://doi.org/10.1109/RE.2019.00025

[14] Fabiano Dalpiaz, Davide Dell'Anna, Fatma Başak Aydemir, and Sercan Çevikol. 2019. explainable-re/RE-2019-Materials v0. https://doi.org/10.5281/zenodo.3309669

[15] Fabiano Dalpiaz, Alessio Ferrari, Xavier Franch, and Cristina Palomares. 2018. Natural language processing for requirements engineering: The best is yet to come. *IEEE Software* 35, 5 (2018), 115–119.

[16] Fabiano Dalpiaz, Ivor van der Schalk, and Garm Lucassen. 2018. Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and NLP. In *Proceedings of the 24th Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ'18)*.

[17] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340. http://www.jstor.org/stable/249008

[18] F. De Bruijn and H. Dekkers. 2010. Ambiguity in natural language software requirements: A case study. In *Proceedings of the International Workshop on Requirements Engineering: Foundation of Software Quality (REFSQ)* (Essen, DE). 233–247.

[19] Cleyton V. C. De Magalhães, Fabio Q. B. da Silva, Ronnie E. S. Santos, and Marcos Suassuna. 2015. Investigations about replication of empirical studies in software engineering: A systematic mapping study. *Information and Software Technology* 64 (2015), 76–101.

[20] Jacek Dąbrowski, Emmanuel Letier, Anna Perini, and Angelo Susi. 2023. Mining and searching app reviews for requirements engineering: Evaluation and replication studies. *Information Systems* 114 (2023), 102181. https://doi.org/10.1016/j.is.2023.102181

[21] Saad Ezzini, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. 2021. taphsir v6. https://doi.org/10.5281/zenodo.6009790

[22] Saad Ezzini, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. 2022. Automated handling of anaphoric ambiguity in requirements: A multi-solution study. In *2022 IEEE/ACM 44th International Conference on Software Engineering*.

[23] Saad Ezzini, Sallam Abualhaija, Chetan Arora, Mehrdad Sabetzadeh, and Lionel C. Briand. 2021. Using domain-specific corpora for improved handling of ambiguity in requirements. In *2021 IEEE/ACM 43rd International Conference on Software Engineering*.

[24] D. Méndez Fernández, Stefan Wagner, Marcos Kalinowski, Michael Felderer, Priscilla Mafra, Antonio Vetrò, Tayana Conte, M.-T. Christiansson, Des Greer, Casper Lassenius, et al. 2017. Naming the pain in requirements engineering. *Empirical Software Engineering* 22, 5 (2017), 2298–2338.

[25] Alessio Ferrari, Felice Dell'Orletta, Andrea Esuli, Vincenzo Gervasi, and Stefania Gnesi. 2017. Natural language requirements processing: A 4D vision. *IEEE Softw.* 34, 6 (2017), 28–35.

[26] Alessio Ferrari, Giorgio Oronzo Spagnolo, and Stefania Gnesi. 2017. PURE: A dataset of public requirements documents. In *2017 IEEE 25th International Requirements Engineering Conference*. 502–505. https://doi.org/10.1109/RE.2017.29

[27] Xavier Franch, Cristina Palomares, Carme Quer, Panagiota Chatzipetrou, and Tony Gorschek. 2023. The state-of-practice in requirements specification: An extended interview study at 12 companies. *Requirements Engineering* (2023). https://doi.org/10.1007/s00766-023-00399-7

[28] Vincenzo Gervasi, Alessio Ferrari, Didar Zowghi, and Paola Spoletini. 2019. Ambiguity in requirements engineering: Towards a unifying framework. In *From Software Engineering to Formal Methods and Tools, and Back*. Springer.

[29] Martin Glinz. 2007. On non-functional requirements. In *15th IEEE International Requirements Engineering Conference (RE '07)*. IEEE, 21–26.

[30] Jesús M. González-Barahona and Gregorio Robles. 2012. On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empirical Software Engineering* 17 (2012), 75–89.

[31] Jesus M. Gonzalez-Barahona and Gregorio Robles. 2023. Revisiting the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Information and Software Technology* 164 (2023), 107318.

[32] Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community expectations for research artifacts and evaluation processes. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 469–480.

[33] David W. Johnson and Frank P. Johnson. 1991. *Joining Together: Group Theory and Group Skills*. Prentice-Hall, Inc.

[34] Natalia Juristo and Sira Vegas. 2011. The role of non-exact replications in software engineering experiments. *Empirical Software Engineering* 16, 3 (2011), 295–324.

[35] Erik Kamsties and Barbara Peach. 2000. Taming ambiguity in natural language requirements. In *Proceedings of the 13th International Conference on Software and Systems Engineering and Applications (ICSSEA'00)*.

[36] Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2676–2686. https://doi.org/10.18653/v1/P18-1249

[37] Nadzeya Kiyavitskaya, Nicola Zeni, Luisa Mich, and Daniel Berry. 2008. Requirements for tools for ambiguity identification and measurement in natural language requirements specifications. *Requirements Engineering* 13, 3 (2008).

[38] Zijad Kurtanović and Walid Maalej. 2017. Automatically classifying functional and non-functional requirements using supervised machine learning. In *2017 IEEE 25th International Requirements Engineering Conference (RE '17)*. IEEE, 490–495.

[39] Zijad Kurtanović and Walid Maalej. 2018. On user rationale in software engineering. *Requirements Engineering* 23, 3 (2018), 357–379.

[40] J. Richard Landis and Gary G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33, 2 (1977).

[41] Feng-Lin Li, Jennifer Horkoff, John Mylopoulos, Renata S. S. Guizzardi, Giancarlo Guizzardi, Alexander Borgida, and Lin Liu. 2014. Non-functional requirements as qualities, with a spice of ontology. In *2014 IEEE 22nd International Requirements Engineering Conference (RE '14)*. IEEE, 293–302.

[42] Daniel Mendez, Daniel Graziotin, Stefan Wagner, and Heidi Seibold. 2020. Open science in software engineering. *Contemporary Empirical Methods in Software Engineering* (2020), 477–501.

[43] Daniel Méndez Fernández, Martin Monperrus, Robert Feldt, and Thomas Zimmermann. 2019. The open science initiative of the Empirical Software Engineering journal. *Empirical Software Engineering* 24 (2019), 1057–1060.

[44] Lloyd Montgomery, Davide Fucci, Abir Bouraffa, Lisa Scholz, and Walid Maalej. 2022. Empirical research on requirements quality: A systematic mapping study. *Requirements Engineering* (2022), 1–27.

[45] Erik Jan Philippo, Werner Heijstek, Bas Kruiswijk, Michel R. V. Chaudron, and Daniel M. Berry. 2013. Requirement ambiguity not as important as expected — results of an empirical evaluation. In *Proceedings of the International Workshop on Requirements Engineering: Foundation of Software Quality (REFSQ '13)* (Essen, DE). 65–79.

[46] Shekoufeh Rahimi, Kevin Charles Lano, and Chenghua Lin. 2022. Requirement formalisation using natural language processing and machine learning: A systematic review. In *International Conference on Model-Based Software and Systems Engineering*. SCITEPRESS Digital Library, 1–8.

[47] Cristina Ribeiro and Daniel Berry. 2020. The prevalence and severity of persistent ambiguity in software requirements specifications: Is a special effort needed to find them? *Science of Computer Programming* 195 (2020), 102472.

[48] Faiz Ali Shah, Kairit Sirts, and Dietmar Pfahl. 2019. Is the SAFE approach too simple for app feature extraction? A replication study. In *Requirements Engineering: Foundation for Software Quality*, Eric Knauss and Michael Goedicke (Eds.). Springer International Publishing, Cham, 21–36.

[49] Martin Shepperd, Nemitari Ajienka, and Steve Counsell. 2018. The role and value of replication in empirical software engineering results. *Information and Software Technology* 99 (2018), 120–132.

[50] Forrest J. Shull, Jeffrey C. Carver, Sira Vegas, and Natalia Juristo. 2008. The role of replications in empirical software engineering. *Empirical Software Engineering* 13, 2 (2008), 211–218.

[51] Roel J. Wieringa. 2014. *Design Science Methodology for Information Systems and Software Engineering*. Springer.
[52] Jonas Winkler and Andreas Vogelsang. 2016. Automatic classification of requirements based on convolutional neural networks. In *24th IEEE International Requirements Engineering Conference, RE 2016, Beijing, China, September 12–16, 2016*. IEEE Computer Society, 39–45. https://doi.org/10.1109/REW.2016.021
[53] Stefan Winter, Christopher S. Timperley, Ben Hermann, Jurgen Cito, Jonathan Bell, Michael Hilton, and Dirk Beyer. 2022. A retrospective study of one decade of artifact evaluations. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 145–156.
[54] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.
[55] Hui Yang, Anne de Roeck, Vincenzo Gervasi, Alistair Willis, and Bashar Nuseibeh. 2011. Analysing anaphoric ambiguity in natural language requirements. *Requirements Engineering* 16, 3 (May 2011), 163. https://doi.org/10.1007/s00766-011-0119-y
[56] Hui Yang, Anne de Roeck, Alistair Willis, and Bashar Nuseibeh. 2010. A methodology for automatic identification of nocuous ambiguity. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling '10)*. Coling 2010 Organizing Committee, Beijing, China, 1218–1226. https://aclanthology.org/C10-1137
[57] Liping Zhao, Waad Alhoshan, Alessio Ferrari, Keletso J. Letsholo, Muideen A. Ajagbe, Erol-Valeriu Chioasca, and Riza T. Batista-Navarro. 2021. Natural language processing for requirements engineering: A systematic mapping study. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–41.
[58] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).