Information and Software Technology 161 (2023) 107252

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Case study identification: A trivial indicator outperforms human classifiers

Austen Rainer [a], Claes Wohlin [b,*]

[a] Queen's University Belfast, 18 Malone Road, Computer Science Building, BT9 5BN, Belfast, Northern Ireland, UK
[b] Blekinge Institute of Technology, SE-371 79, Karlskrona, Sweden

ARTICLE INFO

Keywords:
Case study
Evaluation
Systematic review
Primary study
Smell indicator

ABSTRACT

Context: The definition and term "case study" are not being applied consistently by software engineering researchers. We previously developed a trivial "smell indicator" to help detect the misclassification of primary studies as case studies.
Objective: To evaluate the performance of the indicator.
Methods: We compare the performance of the indicator against human classifiers for three datasets, two datasets comprising classifications by both authors of systematic literature studies and primary studies, and one dataset comprising only primary-study author classifications.
Results: The indicator outperforms the human classifiers for all datasets.
Conclusions: The indicator is successful because human classifiers "fail" to properly classify their own, and others', primary studies. Consequently, reviewers of primary studies and authors of systematic literature studies could use the classifier as a "sanity" check for primary studies. Moreover, authors might use the indicator to double-check how they classified a study, as part of their analysis, and prior to submitting their manuscript for publication. We challenge the research community to both beat the indicator, and to improve its ability to identify true case studies.

1. Introduction

In two recent articles, we demonstrated empirically that the definition and the term "case study" are not applied consistently by software engineering researchers. This problem occurs with primary studies [1] and with Systematic Literature Studies (SLSs, e.g., systematic reviews and systematic mapping studies) that classify primary studies [2]. We show [2] that SLS authors often repeat the incorrect classification of the authors of the primary study but also, and even worse, incorrectly classify primary studies as case studies when the primary-study authors do not classify their papers as case studies. Our two recent articles extend a preliminary study [3], which made a similar finding for primary studies, and complements previous work, by [4,5], that already recognise this problem.

For our research, we chose to use Wohlin's (2021) definition of a case study. Wohlin's definition adds clarifications to Runeson et al.'s (2012) definition, which in turn builds on Yin's (2018) definition. Furthermore, Wohlin's definition distinguishes five components of a case study. These components underpin a checklist that was formulated from the definition [1], and which we use in our assessment. Wohlin's definition is as follows:

> "A case study is an empirical investigation of a case, using multiple data collection methods, to study a contemporary phenomenon in its real-life context, and with the investigator(s) not taking an active role in the case investigated". [3]

Given the problem with the misclassification of case studies, we wanted to find a simple way to *indicate* whether a claimed case study is indeed a case study. Thus, as part of our investigation into how authors classify their own primary studies [1], we proposed and applied a smell indicator to help researchers retrospectively check whether a primary study *claiming* to be a case study is likely to be a case study. We applied that indicator only to the dataset we directly investigated in that study.

We are interested to know how the smell indicator performs against the primary-study authors and the SLS authors we investigated in [2] and the primary-study authors investigated in [3]. Thus, the aims of

[*] Corresponding author.
E-mail address: claes.wohlin@bth.se (C. Wohlin).

https://doi.org/10.1016/j.infsof.2023.107252
Received 28 November 2022; Received in revised form 5 May 2023; Accepted 11 May 2023
Available online 18 May 2023

the current article are to evaluate the indicator, through application to additional datasets, and to discuss the indicator's performance.

The remainder of this article is organised as follows. Section 2 provides an overview to the smell indicator. Section 3 presents our research approach. Section 4 presents our evaluation and limitations. Conclusions, including our discussion of the performance of the indicator, are provided in Section 5.

2. Overview to the indicator

In previous research [1], we introduced and applied a smell indicator to automatically check whether a research study is likely to be a case study. The smell indicator was inspired by the concept of a code smell although, in our case, directed towards the research method "case study".

Before selecting the indicator, the two authors each independently developed a candidate indicator. One indicator comprised a simple counting of the occurrence of the string "interview*", while the second candidate indicator sought to align with the definition of case study. The second indicator required the following: "case stud*" mentioned at least ten times, with "practition*" and "industr*" occurring somewhere in the paper, and with the study not being labelled as something other than case study in the title of the paper, e.g., "empirical study" or "empirical analysis". The sign "*" is used to indicate truncation of any form of the words. The performance of the two indicators was compared. We found that the first indicator was both more powerful in its predictions and simpler in its design; but also "contentious" since it is not based on the definition of case study.

Essentially, the indicator comprises a precondition and a rule: given a primary study already classified as a case study (precondition), then the primary study is indicated to be a case study if the term "interview" occurs at least three times in the article reporting the study (rule).

A consequence of the precondition and rule is that the indicator will indicate that a genuine interview study is a case study if the authors have chosen to classify their primary study as a case study (precondition) and the article contains the term "interview" at least three times (rule). Conversely, the indicator would indicate a genuine case study was *not* a case study, if the respective article does not include interviews as one of the data collection methods, and the word is mentioned at least three times. There are, however, two reasons to suggest this converse consequence will rarely occur. First, [7] emphasises that interviews are one of the most important sources of evidence in a case study. We can therefore expect interviews to very often be conducted, and therefore reported, as part of a genuine case study. Second, in our investigation of primary studies [1], we found only *one* genuine case study [8] in which there were no interviews reported. Nevertheless, the indicator will still misclassify genuine interview studies as case studies, if labelled as a case study by the authors.

3. Research approach

Our evaluation comprises two research questions and three datasets. The three datasets, with links to the articles, are available in Supplements 1–3 in the online supplementary material linked in Appendix A. For the research questions, we ask:

RQ1 Does the smell indicator perform better than SLS authors who have classified primary studies as case studies? We investigate this RQ with two of our three datasets.

RQ2 Does the smell indicator perform better than primary-study authors who have classified their primary studies as case studies? We investigate this RQ using all three datasets.

Two of the datasets are drawn from [2] and one is taken from [3]. The first dataset, from [2], comprises a stratified random sample of 68 primary studies, drawn from nine SLSs. These primary studies were classified as case studies by the SLS authors. The second dataset, also drawn from [2], comprises all 35 primary studies classified as case studies by the authors of one of the nine SLSs. This SLS [9] had the highest number of correct case studies, from the nine SLSs analysed in [2]. The third dataset comprises all 100 primary studies used in [3].

As presented in Section 1, we previously classified all the studies using the case study definition by [3] and, in particular, Wohlin and Rainers's (2022) checklist for case study research. The definitions, our checklist, and our classifications are discussed in more detail in [3], and [2]. We use our classification of a primary study as the correct classification when comparing the performance of the indicator against the authors' classifications.

We performed our analysis in two independent ways. First, we performed our analysis using an automated implementation of the indicator. Then, because there may be challenges parsing the text from PDFs, we complemented our automated indicator with an independent manual analysis of the articles, using Adobe.

We emphasise the intended scope and purpose of the indicator: the indicator should only be applied to primary studies that have already been classified, by primary-study authors or by SLS authors, as a case study. Thus, the indicator is intended as a way of assuring the classification of a study, or for helping an author to proactively sense-check their potential classification.

An inevitable consequence of the indicator's intended scope and purpose is that the datasets will be imbalanced, since all articles investigated are labelled "case study" by the authors, be they the authors of primary studies or SLSs. Due to the imbalanced datasets, it becomes unfeasible to calculate some measures of agreement, e.g., Cohen's kappa. Instead, we use accuracy as the measure of agreement.

We also emphasise that the indicator is not a classifier; for example, it does not take a primary study and classify that study as, for example, a case study, or an experiment, or a survey.

Further details about the research methodology, e.g., concerning the identification of the data sets can be found in [2,3].

4. Results

We present our results in order of the research questions. Section 4.1 presents our evaluation of the classifications of the SLS authors. Section 4.2 presents our evaluation of the classifications by the primary-study authors. Finally, Section 4.3 presents the limitations.

4.1. Evaluating the indicator relative to SLS authors' classifications

Table 1 compares the classifications of the SLS authors against the smell indicator (denoted SI in the tables), for the two datasets taken from [2], and using our classification as the "ground truth". The first data set includes 68 primary studies and the second data set 35 primary studies. Our two independent analyses of the PDF files gave the same result, which we report in the tables below. For both datasets, the smell indicator is more accurate than the classifications of the SLS authors. When there are more studies actually being case studies, the SLS authors' accuracy is closer to the accuracy of the smell indicator.

4.2. Evaluating the indicator relative to primary-study authors' classifications

Table 2 presents the same information as Table 1, but now with the classifications of the primary-study authors. Studies not classified as case studies by the primary-study authors are removed from the analysis, given the precondition of the smell indicator. Hence, the number of studies are reduced to 48 and 31 respectively for the two datasets.

The results in Table 2 are similar to those for SLS authors in Table 1, i.e., the smell indicator performs better than the primary-study authors. However, the accuracy is somewhat higher for the primary-study authors for both datasets in comparison to the SLS authors. Given that

Table 1
Confusion matrices for the classifications by the SLS authors.
Source: Data set from [2].

Confusion matrix		SLS author - 68		SI - 68		SLS author - 35		SI - 35	
		CS	Not CS	CS	Not CS	CS	Not CS	CS	Not CS
Our	CS	11	0	6	5	16	0	14	2
assessment	Not CS	57	0	10	47	19	0	8	11
Accuracy		**0.16**		**0.78**		**0.46**		**0.71**	

Notes :
CS = Case study
SI = Smell indicator

Table 2
Confusion matrices for the classifications by the PS authors.
Source: Data set from [2].

Confusion matrix		PS author - 48		SI-48		PS author - 31		SI - 31	
		CS	Not CS	CS	Not CS	CS	Not CS	CS	Not CS
Our	CS	11	0	6	5	15	0	14	1
assessment	Not CS	37	0	6	31	16	0	7	9
Accuracy		**0.23**		**0.77**		**0.48**		**0.74**	

Table 3
Confusion matrices for the 100 primary studies.
Source: Data set from [3].

Confusion matrix		PS author		Smell indicator	
		CS	Not CS	CS	Not CS
[3]	CS	53	0	42	11
assessment	Not CS	47	0	1	46
Accuracy		**0.53**		**0.88**	

SLSs are intended to synthesise the best evidence from research, it is discouraging that the classifications of the SLS authors are worse than if the SLS authors simply accepted the classifications of the primary-study authors.

Table 3 presents the same kind of information as the previous two tables, but now for the third dataset, taken from [3]. The smell indicator performs strongly in comparison with the primary-study authors.

Earlier we recognised that the datasets are imbalanced. Because the third dataset is fairly evenly balanced in its proportion of case-study to not case-study (a proportion of 53:47) we can use this dataset to gain insights into where the indicator is more and less successful. We observe from Table 3 that 11 genuine case studies are not identified as case studies by the smell indicator and only one study is incorrectly identified as a case study, when it is not actually a case study. This suggests that the indicator is better at indicating primary studies that are *not* case studies but fails to detect about a quarter of true case studies.

4.3. Limitations

As with all research, our research comes with some limitations. First, the articles used in our analysis are taken from [2,3]. Our datasets therefore come with the limitations already recognised in those two articles, where we discuss the limitations in relation to the data sets. Second, we did not perform a quality assessment of the studies included; we treated them as being representative in quality for studies in software engineering. Third, the indicator is intended to be used within software engineering; we do now know how it works in other areas of research. Fourth, we chose to primarily base our assessment on consensus discussions based on individual preparations. Thus, we have not calculated any statistical measure of our agreement when we classified the studies. Finally, the assessment is based on the definition by [3], which is aligned with the definition by [6,7]. Further details concerning the limitations may be found in [2,3].

5. Conclusions

Overall, our evaluation finds that the smell indicator performs better than the respective human classifiers for all datasets evaluated. From a general software engineering research perspective, this result is disappointing: a *trivial* smell indicator outperforms both primary-study authors and SLS authors.

One potential implication is that reviewers of primary studies and authors of systematic literature studies could use the indicator as a quick check on whether a (primary) study is correctly labelled by the authors. Moreover, authors of primary studies might use the indicator to double-check how they classified their study, as part of their analysis, and prior to submitting the manuscript for publication. We also strongly emphasise that the indicator should *not* be used to classify studies as case studies, i.e., using the word "interview" in a study does not make the study a case study. The indicator should *only* be applied to studies *already* labelled as "case study" prior to using the indicator.

The core problem is that researchers in software engineering do not classify their studies in accordance with accepted definitions for different research methods. This was also observed for experiments, by [10].

As noted in Section 1, the problem of classifying case studies has already been emphasised by a number of previous studies [1,3–5]. The contribution of the current article is to demonstrate the problem by comparing the performance of human classifiers against a *trivial* smell indicator.

A counter-intuitive, long-term objective is that we actually want the indicator to *decline* in performance over time relative to researchers classifying their own, and others' work. A relative decline in performance over time might provide a proxy measure for the improvement of research quality, at least in regards to the classification of primary studies as (true) case studies. Thus, we challenge the software engineering research community to *beat* the trivial smell indicator in classifying studies as case studies. In parallel, we suggest that future research seeks to improve the smell indicator's ability to identify *true* case studies. Further research might also develop equivalent indicators for other research methods.

CRediT authorship contribution statement

Austen Rainer: Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Claes Wohlin:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is provided as a supplement to the article

Acknowledgement

Both authors have contributed to all activities related to the submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.infsof.2023.107252.

References

[1] C. Wohlin, A. Rainer, Is it a case study?—A critical analysis and guidance, J. Syst. Softw. 192 (2022) 111395.

[2] A. Rainer, C. Wohlin, Reporting case studies in systematic literature studies – an evidential problem, 2023, A copy may be obtained from the authors. (Submitted for publication).

[3] C. Wohlin, Case study research in software engineering—It is a case, and it is a study, but is it a case study? Inf. Softw. Technol. 133 (2021) 106514.

[4] C. Zannier, G. Melnik, F. Maurer, On the success of empirical studies in the international conference on software engineering, in: Proceedings International Conference on Software Engineering, 2006, pp. 341–350.

[5] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empir. Softw. Eng. 14 (2) (2009) 131.

[6] P. Runeson, M. Höst, A. Rainer, B. Regnell, Case Study Research in Software Engineering: Guidelines and Examples, Wiley Publishing, 2012.

[7] R.K. Yin, Case Study Research and Applications: Design and Methods, sixth ed., Sage Publications, 2018.

[8] A. Mockus, R.T. Fielding, J.D. Herbsleb, Two case studies of open source software development: Apache and Mozilla, ACM Trans. Softw. Eng. Methodol. 11 (3) (2002) 309–346.

[9] A.E. Chacón-Luna, A.M. Gutiérrez, J.A. Galindo, D. Benavides, Empirical software product line engineering: a systematic literature review, Inf. Softw. Technol. 128 (2020) 106389.

[10] C. Ayala, B. Turhan, X. Franch, N. Juristo, Use and misuse of the term "experiment" in mining software repositories research, IEEE Trans. Softw. Eng. 48 (11) (2022) 4229–4248.