Agile Effort Estimation: Have We Solved the Problem Yet? Insights From A Replication Study (Supplementary Material)

Vali Tawosi, Rebecca Moussa, Federica Sarro
{*vali.tawosi, rebecca.moussa.18, f.sarro*}*@ucl.ac.uk*

Abstract

This document is supplementary to the paper entitled "Agile Effort Estimation: Have We Solved the Problem Yet? Insights From A Replication Study?", which is accepted for publication in the IEEE Journal of Transactions on Software Engineering. Here, we report supplementary results and additional information about the large dataset of issues used in the evaluation in our study (i.e., the Tawosi dataset).

1 Supplementary Results

We refer the reader to our paper [1] for a discussion on the results with respect to the MAE values. Here, we present and analyse the results with respect to the median Absolute Error (MdAE) and Standard Accuracy (SA) values.

1.1 RQ1.1

Table 1 shows the results obtained by Deep-SE and the baseline techniques on the Choet dataset in our replication. Specifically, column "Rep" shows the results of replication with the transformation applied on the SP distribution (more information regarding the transformation can be found in [1]). Results by applying the transformation only on the training set are shown under "CutTrain" column, and results without applying the transformation (using the SP distribution as-is) are presented under the "!Cut" column. The results reported by the original study [2] are reported under the "Orig" column for reference.

As we can see, all the SA values achieved by the techniques in Table 1 are positive, which means that all the techniques outperform the Random Guessing baseline in terms of MAE. However, a noticeable decrease is observable between the SA values achieved in the replication study and the original study for most of the cases. This can be due to a possibly different approach taken by the original study to perform Random Guessing, which directly affects the computation of the SA.

Choetkiertikul et al. [2] state that they compute Random Guessing as recommended in the literature [3–6]. Specifically, the explanation in Section 5.3 of the original study defines MAE_{rguess} (MAE_{p_0} in Equation (1) in our study) as the MAE of a large number (e.g., 1000 runs) of random

guesses. We followed the same definition, in our replication study, however we noticed that the MAE_{rguess} values we computed are different (i.e., they are generally much lower than those used in the original study) thus resulting in a smaller SA in most of the cases. While we do expect Random Guessing to be different to some extent, given its stochastic nature, the differences observed between our results and the ones in the original study are simply too large: they range from 0.18% to 130.36% (with a mean difference of 41.19%) when the data is not transformed, and from 0.04% to 52.24% (with a mean difference of 22.86%) when a transformation is applied on the data. Based on both the inspection of the code and the replies of the authors, we could not find a definite explanation for such a large difference in the results. In fact, we could not find the code to compute random guessing in the replication package[1], and the authors confirmed that some calculations, such as SA, were indeed performed with different tools (e.g., MATLAB, Excel) and not maintained within the repository. The replication package only contains a file named random.txt for each of the projects, which contains one SP value per issue. This may suggest that only one value per random guess had been produced instead of 1,000 runs, however, the MAE of these values do not match the MAE_{rguess} used in the original paper.

Considering MdAE, Deep-SE ("Rep") achieved a lower (better) MdAE than each of the Mean and Median baselines in 13 cases (81%) out of the 16 under study. Deep-SE with transformation applied on the train set ("CutTrain") performs similar to "Rep" (i.e., Deep-SE ("CutTrain") achieved a lower MdAE than each of the Mean and Median baselines in 13 out of the 16 cases (81%)). However, when no transformation is applied ("!Cut"), Deep-SE achieves lower MdAE than both Mean and Median baselines in all 16 cases. Deep-SE ("Rep") produced a higher (worse) MdAE than Deep-SE ("Orig") in 11 out of the 16 cases (69%).

$$SA = \left(1 - \frac{MAE_{p_i}}{MAE_{p_0}}\right) \times 100 \tag{1}$$

where MAE_{p_i} is the MAE of the approach p_i being evaluated and MAE_{p_0} is the MAE of a large number (usually $1,000$ runs) of *random guesses*.

1.2 RQ1.2

Table 2 shows the results of Deep-SE, Mean and Median baselines on the Tawosi dataset. As we can see, Deep-SE consistently achieves a positive SA, meaning that it outperforms the Random Guessing baseline in all 26 cases.

With respect to MdAE, Deep-SE achieves better results (lower MdAE) than both the Mean and Medina baselines in 19 out of 26 cases (73%). Furthermore, compared to the baselines individually, Deep-SE outperforms the Mean and Median, with respect to MdAE values achieved, in 21 (81%) and 22 (85%) cases, respectively.

1.3 RQ2.1

Table 3 shows the results we obtained in our replication study using Deep-SE ("Rep") and TF/IDF-SVM ("Rep"), together with the results obtained by the original study (Deep-SE (Orig) and TF/IDF-SVM (Orig), shaded in grey) on the Porru's dataset. The results of the Mean and Median baselines are also provided.

[1]We could determine the MAE_{rguess} used in the original study by reversing its computation given that the original study reports both SA and MAE.

Again, all the SA values are positive, which means that all the approaches outperformed the Random Guessing baseline. Comparing Deep-SE to the other techniques, the former produced a lower MdAE than each of TF/IDF-SVM, Mean and Median baselines in 6 out of 8 cases (75%). TF/IDF-SVM reaches to the same MdAE achieved by the Median baseline in 7 out of 8 cases (88%), where the XD project is the exception.

1.4 RQ3.1

Table 4 shows the results of cross-project estimation. In this table, all the SA values are positive, meaning that all the techniques are able to outperform the Random Guessing baseline. However, the SA value is very low for three cases among the cross-repository experiments (0.98%, 6.18%, and 7.37%). Comparing Deep-SE's performance, with respect to MdAE measure, against the Mean and Median baseline, we observe that Deep-SE achieves a lower MdAE than the Mean and Median in 9 and 8 cases out of 16 cases.

1.5 RQ3.2

Considering the cross-project estimations on the Tawosi dataset, presented in Table 5, we see that Deep-SE outperforms the Mean baseline in 5 and the Median baseline in 3 cases out of 5. Furthermore, the SA values show that all three techniques outperform the Random Guessing in all 5 cases.

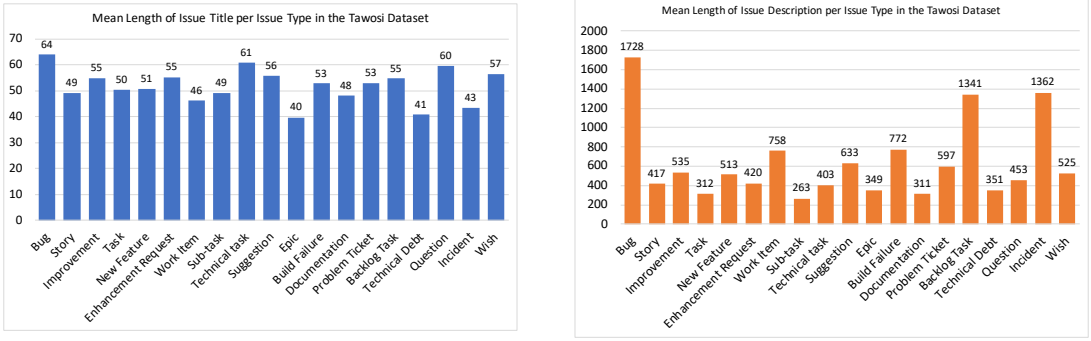
1.6 RQ4

Table 6 shows the results of Deep-SE and the baseline techniques with augmented training sets. As we can see, with regards to MdAE values, Deep-SE outperforms the Mean baseline in 15 and the Median baseline in 14 out of 18 cases (78%). All positive values for SA show that Deep-SE outperforms the Random Guessing baseline.

1.7 RQ5

Tables 1 and 2 reports the results of the comparisons of Deep-SE with and without pre-training (Deep-SE vs Deep-SE!pre-train) on the Choet dataset and the Tawosi dataset, respectively. Overall, we can observe that Deep-SE achieves a lower MdAE than Deep-SE!pre-train in 21 cases out of 42 cases (50%), a higher MdAE in 20 (48%) cases and an equal MdAE on the remaining TESB project from the Tawosi dataset. With regards to SA values, Deep-SE!pre-train achieved a positive value, thus outperforming the Random Guessing baseline, in all the cases except for the STL project in the Tawosi dataset, where an SA value of -0.33% is achieved. It is worth noting that the SA value for this project using Deep-SE with pre-training is also low (i.e., 1.91%).

According to the comparison between the running time of Deep-SE and Deep-SE!pre-train in the paper, the two variants take almost the same amount of time to run. Here, we report running time for Deep-SE and TF/IDF-SVM on the Tawosi dataset in Table 7 for completeness. As we can see, in total, TF/IDF-SVM runs in less than 1% of the time taken by Deep-SE.



(a) (b)

Figure 1: Mean length of issue title (a) and description (b) in the Tawosi dataset per issue type (in characters). Issue types are ordered by their frequency in descending order from left to right.

2 Dataset Description

In this study, in addition to the datasets used by Choetkiertikul et al. [2], and Porru et al. [7], we used the Tawosi dataset and made it publicly available at [8]. This dataset is sampled from the larger TAWOS dataset which includes more than half a million issue reports from 44 open-source projects [9]. Our sampled dataset contains a total of 31,960 issues with story points from 26 projects belonging to 13 different repositories:

One Project from **Apache**:

1. **Apache Mesos** (MESOS) is an open-source project to manage computer clusters. It was developed in C++ language at the University of California, Berkeley.

Seven Project from **Appcelerator**:

2. **Alloy Framework** (ALOY) is an Apache-licensed model–view–controller application framework built on top of Titanium which provides a simple model for separating the application user interface, business logic and data models.
3. **Aptana Studio** (APSTUD) is an open-source integrated development environment (IDE) for building web applications.
4. **Appcelerator Command-Line Interface** (CLI) Command-Line Interface is provided by Appcelerator to check and configure environment setup, create, and build applications.
5. **Appcelerator DAEMON** (DAEMON) The Appcelerator Daemon is a server that runs on a developer's computer and hosts services which power the tooling for Axway products such as Axway Titanium SDK.
6. **Titanium Mobile Platform** (TIDOC) Titanium Mobile is a mature platform for developers to build completely native cross-platform mobile applications.
7. **The Titanium SDK** (TIMOB) is the software development kit for Titanium platform.
8. **Appcelerator Studio** (TISTUD) Appcelerator Studio is an eclipse based IDE that provides a single, extensible environment to rapidly build, test, package, and publish mobile apps across multiple devices and OSs.

Three Projects from **Atlassian**:

9. **Atlassian Clover** (CLOV) Clover is a Java code coverage analysis utility bought and further developed by Atlassian.

10, and 11. **Atlassian Confluence Cloud and Server** (CONFCLOUD and CONFSERVER) Confluence is a knowledge sharing tool that helps teams create and share content.

One project from **DNNSoftware**

12. **DNN Platform** (DNN) DNN is a web content management system and web application framework based on the .NET Framework. The DNN Platform Edition is open source and written in C#.

One project from **DuraSpace**:

13. **Lyrasis Dura Cloud** (DURACLOUD) DuraCloud is a hosted service from LYRASIS that lets you control where and how your content is preserved in the cloud.

Two projects from **Hyperledger**:

14. **Hyperledger Fabric** (FAB) Hyperledger Fabric is intended as a foundation for developing applications or solutions with a modular architecture. Hyperledger Fabric allows components, such as consensus and membership services, to be plug-and-play. Its modular and versatile design satisfies a broad range of industry use cases. It offers a unique approach to consensus that enables performance at scale while preserving privacy.

15. **Hyperledger Sawtooth** (STL) Hyperledger Sawtooth offers a flexible and modular architecture separates the core system from the application domain.

One project from **Lsstcorp**:

16. **Lsstcorp Data management** (DM) Data Management is responsible for creating the software, services and systems which will be used to produce Rubin Observatory's data products.

Three projects from **MongoDB**: MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema.

17. **MongoDB Compass** (COMPASS) MongoDB Compass provides quick visualization of the structure of data in the database, and perform ad hoc queries – all with zero knowledge of MongoDB's query language.

18. **MongoDB Core Server** (SERVER) MongoDB Enterprise Server is the commercial edition of MongoDB, available as part of the MongoDB Enterprise Advanced subscription.

19. **MongoDB Evergreen** (EVG) MongoDB Evergreen is a continuous integration system, customise build for MongoDB ecosystem.

One project from **Moodle**:

20. **Moodle** (MDL) Moodle is a free and open-source learning management system (LMS) written in PHP and distributed under the GNU General Public License.

One project from **Mulesoft**:

21. **Mule** (MULE) Mule is a lightweight enterprise service bus (ESB) and integration framework provided by MuleSoft. The platform is Java-based, but can broker interactions between other

platforms such as .NET using web services or sockets.

One project from **Sonatype**:

22. **Sonatype Nexus** (NEXUS) Nexus is a repository manager. It allows developers to proxy, collect, and manage their dependencies.

One project from **Spring**:

23. **Spring XD** (XD) Spring XD is a unified, distributed, and extensible service for data ingestion, real time analytics, batch processing, and data export.

Three projects from **Talendforge**:

24. **Talend Data Preparation** (TDP) Talend Data Preparation is a self-service application that enables information workers to prepare data for analysis and other data-driven tasks.

25. **Talend Data Quality** (TDQ) Talend's Data Quality profiles, cleans, and masks data in any format or size to deliver data you can trust for the insights you need.

26. **Talend ESB** (TESB) Talend ESB is a reliable and scalable enterprise service bus (ESB) that lets development teams manage integration projects in a holistic manner, combining integration of applications and data management in complex and heterogeneous computing environments.

The Tawosi dataset includes a variety of issue types. Tables 8, 9, and 10 present the descriptive statistics of the length of issue title and description per each issue type in each project. Figure 1 shows the mean length of issue title and description per issue type in the whole dataset.

References

[1] V. Tawosi, R. Moussa, and F. Sarro, "Agile effort estimation: Have we solved the problem yet? insights from a replication study," *IEEE Transactions on Software Engineering (TSE)*, 2022.

[2] M. Choetkiertikul, H. K. Dam, T. Tran, T. Pham, A. Ghose, and T. Menzies, "A deep learning model for estimating story points," *IEEE Transactions on Software Engineering (TSE)*, vol. 45, no. 7, pp. 637–656, 2019.

[3] M. Shepperd and S. MacDonell, "Evaluating prediction systems in software project estimation," *Information and Software Technology*, vol. 54, no. 8, pp. 820–827, 2012.

[4] W. B. Langdon, J. Dolado, F. Sarro, and M. Harman, "Exact mean absolute error of baseline predictor, marp0," *Information and Software Technology*, vol. 73, pp. 16–18, 2016.

[5] F. Sarro, A. Petrozziello, and M. Harman, "Multi-objective software effort estimation," in *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*. IEEE, 2016, pp. 619–630.

[6] V. Tawosi, F. Sarro, A. Petrozziello, and M. Harman, "Multi-objective software effort estimation: A replication study," *IEEE Transactions on Software Engineering (TSE)*, 2021.

[7] S. Porru, A. Murgia, S. Demeyer, M. Marchesi, and R. Tonelli, "Estimating story points from issue reports," in *PROMISE*, 2016, pp. 1–10.

[8] "Deep-SE, fixed Python source code, TF/IDF-SVM Python source code, and the datasets used in this study." [Online]. Available: https://github.com/SOLAR-group/AgileEffortEstimation

[9] V. Tawosi, A. Al-Subaihin, R. Moussa, and F. Sarro, "A versatile dataset of agile open source software projects," in *Proceedings of 19th International Conference on Mining Software Repositories*. ACM, 2022.

Table 1: RQ1.1 and RQ5. Results obtained for the Choet dataset in RQ1.1.: The column "Rep" shows the replication results, the column "Orig" presents the original study results [2], both obtained by using Deep-SE with the transformed SPs as done in the original study. The column "CutTrain" shows the results achieved by applying the transformation only on the training set, while the column "!Cut" shows the results of our replication without transforming the SPs. We also include in this table the results for RQ5 "Deep-SE!pre-train", which investigates Deep-SE without pre-training its lower layers (i.e., word embedding and LSTM). The best results per project are highlighted in bold.

Project	Method	MAE				MdAE				SA			
		Rep	CutTrain	!Cut	Orig	Rep	CutTrain	!Cut	Orig	Rep	CutTrain	!Cut	Orig
MESOS	Deep-SE	1.05	**1.15**	**1.12**	**1.02**	**0.69**	**0.71**	**0.73**	**0.74**	32.84	**30.79**	**46.72**	**59.84**
	Deep-SE!pre-train	**1.02**				0.81				**34.74**			
	Mean	1.11	1.22	1.41	1.64	0.85	0.85	1.78	1.31	28.91	27.11	33.18	35.61
	Median	1.11	1.22	1.22	1.73	1.00	1.00	2.00	1.00	28.83	27.03	42.30	32.01
USERGRID	Deep-SE	1.06	1.18	1.18	**1.03**	0.84	0.87	**0.80**	0.87	21.91	20.24	23.25	**52.66**
	Deep-SE!pre-train	1.11				0.92				18.06			
	Mean	1.09	1.21	1.19	1.48	**0.78**	**0.78**	1.23	**0.84**	19.94	18.29	22.43	32.13
	Median	**1.03**	**1.15**	**1.15**	1.60	1.00	1.00	1.00	1.00	**23.95**	**21.97**	**24.97**	26.29
TISTUD	Deep-SE	1.41	1.43	1.42	**1.36**	1.13	1.13	**0.58**	1.04	40.85	40.64	53.49	**60.26**
	Deep-SE!pre-train	1.42				1.11				40.27			
	Mean	1.52	1.55	1.91	2.08	0.44	0.44	1.52	1.09	36.23	35.85	37.52	39.02
	Median	**1.28**	**1.30**	**1.30**	1.84	**0.00**	**0.00**	1.00	**0.00**	**46.47**	**45.97**	**57.41**	46.17
APSTUD	Deep-SE	3.57	4.09	4.14	**2.71**	3.39	3.41	**2.52**	3.27	19.33	17.53	26.74	**42.58**
	Deep-SE!pre-train	3.15				3.05				28.79			
	Mean	**2.95**	**3.48**	**3.59**	3.15	**2.12**	**2.12**	3.46	**2.86**	**33.33**	**29.76**	**36.53**	33.30
	Median	3.08	3.61	3.61	3.71	3.00	3.00	4.00	3.00	30.48	27.22	36.18	21.54
TIMOB	Deep-SE	2.10	2.19	2.09	**1.97**	1.78	**1.80**	**1.34**	1.90	44.16	43.02	54.49	**55.92**
	Deep-SE!pre-train	2.04				**1.64**				45.62			
	Mean	2.53	2.62	3.02	3.05	1.89	1.89	1.97	**1.87**	32.66	31.89	34.29	31.59
	Median	**1.94**	**2.04**	**2.04**	2.47	2.00	2.00	2.00	2.00	**48.18**	**47.04**	**55.71**	44.65
BAM	Deep-SE	0.80	0.80	0.81	**0.74**	**0.73**	**0.72**	**0.61**	**0.60**	40.53	40.86	51.78	**71.24**
	Deep-SE!pre-train	0.77				0.83				43.05			
	Mean	1.03	1.03	1.22	1.75	1.00	1.00	1.31	1.00	24.07	24.07	27.99	32.11
	Median	**0.75**	**0.75**	**0.75**	1.32	1.00	1.00	1.00	1.00	**44.24**	**44.24**	**55.34**	48.72
CLOV	Deep-SE	2.46	3.75	**3.39**	**2.11**	0.89	**0.89**	**0.80**	**0.93**	34.03	25.27	**39.41**	**50.45**
	Deep-SE!pre-train	**2.37**				**0.73**				**36.83**			
	Mean	2.97	4.26	4.57	3.49	3.28	3.28	3.06	3.06	20.40	15.15	18.46	17.84
	Median	2.42	**3.71**	3.71	2.84	2.00	2.00	2.00	2.00	35.16	**26.12**	33.77	33.33
JSWSERVER	Deep-SE	**1.57**	**1.77**	**1.70**	**1.38**	**1.01**	**1.01**	**1.09**	**1.05**	**39.45**	**36.59**	**50.08**	**59.52**
	Deep-SE!pre-train	1.58				1.43				39.06			
	Mean	1.86	2.07	2.40	2.48	1.46	1.46	2.15	2.11	28.22	26.12	29.61	27.06
	Median	2.10	2.31	2.31	2.93	2.00	2.00	2.00	2.00	18.99	17.57	32.40	13.88
DURACLOUD	Deep-SE	**0.64**	**0.71**	**0.82**	**0.68**	0.46	**0.55**	**0.53**	**0.60**	**40.38**	**43.20**	**46.97**	**69.92**
	Deep-SE!pre-train	0.68				**0.41**				36.77			
	Mean	0.73	0.82	1.00	1.30	0.93	1.06	1.14	1.35	32.13	34.32	35.17	42.88
	Median	0.76	0.82	0.82	0.73	1.00	1.00	1.00	1.00	28.92	34.32	46.78	68.08
DM	Deep-SE	**3.70**	**5.88**	**5.86**	**3.77**	**2.09**	**2.29**	**2.22**	**2.19**	**43.49**	**36.10**	**50.87**	**47.87**
	Deep-SE!pre-train	3.91				2.42				40.27			
	Mean	4.89	7.14	8.66	5.29	4.77	4.63	4.55	7.51	25.30	22.32	27.35	26.85
	Median	4.28	6.19	6.19	4.82	3.00	3.00	3.00	3.00	34.70	32.72	48.09	33.38
MDL	Deep-SE	6.89	6.89	7.89	**5.97**	5.31	5.31	**4.93**	6.60	42.72	42.72	48.92	**50.29**
	Deep-SE!pre-train	8.05				6.88				32.98			
	Mean	10.19	10.19	12.63	10.90	10.89	10.89	12.11	14.29	15.34	15.34	18.21	9.16
	Median	**6.59**	**6.59**	**6.59**	7.18	**5.00**	**5.00**	6.00	**5.00**	**45.21**	**45.21**	**57.32**	40.16
MULE	Deep-SE	2.26	2.53	2.59	**2.18**	2.26	**2.42**	**1.96**	**2.32**	23.50	21.56	29.14	**40.09**
	Deep-SE!pre-train	2.30				**2.13**				22.48			
	Mean	2.22	2.49	2.60	2.59	2.80	2.80	2.22	2.73	24.90	22.84	28.72	28.82
	Median	**2.21**	**2.47**	**2.47**	2.69	3.00	3.00	2.00	3.00	**25.36**	**23.26**	**32.29**	26.07
MULESTUDIO	Deep-SE	**3.12**	**3.66**	3.67	**3.23**	**1.33**	**2.46**	**2.26**	2.48	**23.79**	**15.20**	24.12	**17.17**
	Deep-SE!pre-train	3.24				2.39				20.82			
	Mean	3.22	3.70	3.74	3.34	1.51	3.08	2.80	**2.34**	21.32	14.24	22.65	14.21
	Median	3.18	3.66	**3.66**	3.30	2.00	3.00	3.00	3.00	22.48	15.07	**24.32**	15.42
XD	Deep-SE	1.66	**1.63**	**1.70**	**1.63**	1.33	**1.24**	**1.31**	**1.36**	31.81	**33.91**	**39.96**	**46.82**
	Deep-SE!pre-train	**1.58**				**1.32**				**35.09**			
	Mean	1.88	1.91	2.05	2.27	1.51	1.51	2.53	1.87	22.97	22.65	27.59	26.00
	Median	1.68	1.71	1.71	2.07	2.00	2.00	2.00	2.00	31.09	30.66	39.55	32.55
TDQ	Deep-SE	**2.88**	**2.90**	3.61	**2.97**	**2.64**	**2.47**	**2.92**	3.72	**37.08**	**38.83**	30.07	**48.28**
	Deep-SE!pre-train	2.94				2.85				35.94			
	Mean	4.08	4.25	4.56	4.81	4.28	4.28	5.08	4.72	10.85	10.47	11.69	16.18
	Median	3.15	3.31	**3.31**	3.87	3.00	3.00	4.00	**3.00**	31.22	30.13	**35.85**	32.43
TESB	Deep-SE	**0.61**	**0.85**	**0.90**	**0.64**	**0.56**	**0.54**	**0.59**	**0.59**	**33.81**	**36.56**	**36.19**	**69.67**
	Deep-SE!pre-train	0.63				0.60				31.63			
	Mean	0.71	1.00	1.04	1.14	0.94	1.17	0.91	1.25	22.76	24.97	26.47	45.86
	Median	0.70	0.92	0.92	1.16	1.00	1.00	1.00	1.00	23.70	30.96	34.86	44.44

Table 2: RQs 1.2, 2.2, and 5. Results of Deep-SE, TF/IDF-SVM, and baseline estimators (Mean and Median) on the Tawosi dataset. The best values per project are highlighted in bold.

Project	Method	MAE	MdAE	SA	Project	Method	MAE	MdAE	SA	Project	Method	MAE	MdAE	SA
MESOS	Deep-SE	1.34	1.12	34.07	CONFCLOUD	Deep-SE	1.48	**0.93**	33.89	EVG	Deep-SE	0.63	**0.54**	**19.39**
	Deep-SE!pre-train	1.43	1.06	29.88		Deep-SE!pre-train	1.44	1.14	36.03		Deep-SE!pre-train	**0.62**	0.63	0.48
	TF/IDF-SVM	**1.34**	**1.00**	**34.38**		TF/IDF-SVM	**1.33**	1.00	40.86		TF/IDF-SVM	0.69	1.00	10.67
	Mean	1.37	1.08	32.72		Mean	1.49	1.23	33.65		Mean	0.68	0.56	12.99
	Median	**1.34**	**1.00**	**34.38**		Median	**1.33**	1.00	**40.87**		Median	0.69	1.00	10.68
ALOY	Deep-SE	1.51	**1.28**	39.67	CONFSERVER	Deep-SE	**0.91**	**0.64**	**52.28**	MDL	Deep-SE	**3.55**	**2.77**	**76.00**
	Deep-SE!pre-train	1.71	1.67	31.61		Deep-SE!pre-train	0.96	0.67	49.48		Deep-SE!pre-train	5.08	4.30	65.60
	TF/IDF-SVM	**1.44**	2.00	**42.53**		TF/IDF-SVM	0.96	1.00	49.64		TF/IDF-SVM	6.31	7.00	57.30
	Mean	2.23	2.17	10.84		Mean	1.35	1.45	29.17		Mean	14.54	15.23	1.58
	Median	**1.44**	2.00	**42.53**		Median	0.96	1.00	49.64		Median	6.31	7.00	57.30
TISTUD	Deep-SE	1.63	1.38	48.08	DNN	Deep-SE	0.72	0.69	41.69	MULE	Deep-SE	2.24	1.68	37.95
	Deep-SE!pre-train	1.68	**1.28**	46.67		Deep-SE!pre-train	0.72	**0.59**	41.74		Deep-SE!pre-train	**2.22**	**1.56**	**38.46**
	TF/IDF-SVM	**1.51**	2.00	**51.89**		TF/IDF-SVM	0.79	1.00	36.13		TF/IDF-SVM	3.58	2.00	0.81
	Mean	2.01	2.16	35.93		Mean	0.80	0.88	35.29		Mean	2.79	3.18	22.68
	Median	**1.51**	2.00	**51.89**		Median	**0.71**	1.00	**42.60**		Median	2.24	2.00	38.05
APSTUD	Deep-SE	4.31	2.70	27.37	DURACLOUD	Deep-SE	0.68	**0.58**	39.90	NEXUS	Deep-SE	1.08	0.88	26.56
	Deep-SE!pre-train	4.15	3.00	30.15		Deep-SE!pre-train	0.74	0.87	34.60		Deep-SE!pre-train	**1.05**	0.77	**28.98**
	TF/IDF-SVM	**3.99**	3.00	**32.81**		TF/IDF-SVM	0.68	1.00	39.94		TF/IDF-SVM	1.17	1.00	20.68
	Mean	4.00	**2.49**	32.72		Mean	**0.67**	0.85	**41.13**		Mean	1.11	**0.58**	24.69
	Median	**3.99**	3.00	**32.81**		Median	0.68	1.00	39.94		Median	1.17	1.00	20.68
CLI	Deep-SE	1.76	1.30	33.44	FAB	Deep-SE	0.86	0.71	61.06	XD	Deep-SE	1.45	1.16	43.06
	Deep-SE!pre-train	**1.58**	**1.22**	**40.14**		Deep-SE!pre-train	0.75	**0.60**	65.93		Deep-SE!pre-train	**1.42**	**0.94**	**44.14**
	TF/IDF-SVM	2.98	3.00	-12.84		TF/IDF-SVM	1.10	1.00	50.31		TF/IDF-SVM	2.01	2.00	20.82
	Mean	2.14	2.61	18.93		Mean	1.19	1.10	46.21		Mean	1.65	1.72	34.89
	Median	1.77	2.00	33.04		Median	**0.67**	1.00	**69.76**		Median	1.55	1.00	39.05
DAEMON	Deep-SE	3.29	**2.00**	20.55	STL	Deep-SE	1.18	1.12	1.91	TDP	Deep-SE	0.99	0.81	37.69
	Deep-SE!pre-train	3.00	2.60	27.53		Deep-SE!pre-train	1.20	1.09	-0.33		Deep-SE!pre-train	**0.98**	**0.78**	**38.46**
	TF/IDF-SVM	**2.74**	3.00	**33.81**		TF/IDF-SVM	**0.84**	**0.00**	**30.12**		TF/IDF-SVM	0.99	1.00	37.74
	Mean	2.75	2.75	33.53		Mean	0.97	1.02	19.32		Mean	1.17	1.38	26.26
	Median	**2.74**	3.00	**33.81**		Median	0.95	1.00	20.41		Median	0.99	1.00	37.74
TIDOC	Deep-SE	**2.72**	1.19	**25.35**	DM	Deep-SE	1.61	0.89	52.41	TDQ	Deep-SE	**2.47**	**2.23**	**49.14**
	Deep-SE!pre-train	3.26	2.38	10.41		Deep-SE!pre-train	1.65	**0.79**	50.93		Deep-SE!pre-train	2.92	2.90	39.91
	TF/IDF-SVM	3.03	**1.00**	16.69		TF/IDF-SVM	**1.49**	1.00	**55.71**		TF/IDF-SVM	5.05	5.00	-3.95
	Mean	2.99	2.59	18.00		Mean	2.60	2.43	22.83		Mean	4.20	3.82	13.65
	Median	2.77	**1.00**	24.03		Median	1.61	1.00	52.19		Median	2.88	3.00	40.72
TIMOB	Deep-SE	**2.41**	**1.81**	**33.90**	COMPASS	Deep-SE	1.63	1.34	15.25	TESB	Deep-SE	1.15	0.73	21.36
	Deep-SE!pre-train	2.49	1.85	31.77		Deep-SE!pre-train	1.66	**1.33**	13.76		Deep-SE!pre-train	1.09	**0.73**	25.40
	TF/IDF-SVM	2.53	2.00	30.70		TF/IDF-SVM	**1.38**	2.00	**28.54**		TF/IDF-SVM	**0.97**	1.00	**33.95**
	Mean	2.55	1.81	30.23		Mean	1.48	1.63	23.05		Mean	0.99	0.99	32.71
	Median	2.53	2.00	30.70		Median	**1.38**	2.00	28.54		Median	0.98	1.00	33.02
CLOV	Deep-SE	**3.78**	1.05	**47.73**	SERVER	Deep-SE	0.89	0.71	57.60					
	Deep-SE!pre-train	3.89	1.48	46.33		Deep-SE!pre-train	0.87	**0.65**	58.82					
	TF/IDF-SVM	4.04	**1.00**	44.15		TF/IDF-SVM	0.93	1.00	55.88					
	Mean	5.93	5.30	18.06		Mean	1.56	1.86	25.99					
	Median	4.01	2.00	44.55		Median	**0.85**	1.00	**59.47**					

Table 3: RQ2.1. Results of the Deep-SE and TF/IDF-SVM replication (Rep), original study [2] (Orig), and the baselines on the Porru dataset. The best results per project (among all methods but Deep-SE (Orig)) are highlighted in bold.

Project	Method	MAE	MdAE	SA	Project	Method	MAE	MdAE	SA
	Deep-SE (Orig)	1.44	—[a]	93.67[b]		Deep-SE (Orig)	2.32	—	40.71
	TF/IDF-SVM (Orig)	1.76	—	92.27		TF/IDF-SVM (Orig)	3.37	—	13.77
	Deep-SE (Rep)	7.36	**1.70**	67.65		Deep-SE (Rep)	3.27	**1.76**	16.32
TIMOB	TF/IDF-SVM (Rep)	**1.76**	2.00	**92.27**	MULE	TF/IDF-SVM (Rep)	3.37	3.00	13.77
	Mean	20.08	19.76	11.75		Mean	3.22	3.98	17.77
	Median	**1.76**	2.00	**92.27**		Median	**3.07**	3.00	**21.61**
	Deep-SE (Orig)	1.04	—	62.04		Deep-SE (Orig)	1.00	—	52.91
	TF/IDF-SVM (Orig)	1.28	—	53.44		TF/IDF-SVM (Orig)	1.86	—	12.54
	Deep-SE (Rep)	1.36	0.87	50.23		Deep-SE (Rep)	**1.23**	**0.78**	**41.91**
TISTUD	TF/IDF-SVM (Rep)	**1.28**	**0.00**	**53.44**	XD	TF/IDF-SVM (Rep)	1.86	2.00	12.54
	Mean	1.87	0.99	31.66		Mean	1.24	0.79	41.38
	Median	**1.28**	**0.00**	**53.44**		Median	1.34	1.00	36.86
	Deep-SE (Orig)	2.67	—	58.54		Deep-SE (Orig)	0.47	—	59.97
	TF/IDF-SVM (Orig)	5.69	—	11.58		TF/IDF-SVM (Orig)	1.08	—	8.01
	Deep-SE (Rep)	**5.52**	3.34	**14.31**		Deep-SE (Rep)	**0.69**	0.42	**40.93**
APSTUD	TF/IDF-SVM (Rep)	5.69	3.00	11.58	DNN	TF/IDF-SVM (Rep)	1.08	1.00	8.01
	Mean	5.59	**2.29**	13.15		Mean	0.72	**0.25**	39.00
	Median	5.69	3.00	11.58		Median	1.08	1.00	8.01
	Deep-SE (Orig)	0.76	—	60.91		Deep-SE (Orig)	0.21	—	73.97
	TF/IDF-SVM (Orig)	1.23	—	36.59		TF/IDF-SVM (Orig)	0.39	—	51.87
	Deep-SE (Rep)	1.08	**0.96**	44.34		Deep-SE (Rep)	**0.30**	**0.23**	**62.46**
MESOS	TF/IDF-SVM (Rep)	1.23	1.00	36.59	NEXUS	TF/IDF-SVM (Rep)	0.39	0.50	51.87
	Mean	1.24	1.08	36.15		Mean	0.72	0.86	11.16
	Median	**0.84**	1.00	**57.02**		Median	0.39	0.50	51.88

[a]The original study did not report MdAE for Deep-SE and TF/IDF-SVM.
[b]We used the MAE values reported in the original study to compute SA for Deep-SE (Orig) and TF/IDF-SVM (Orig).

Table 4: RQ3.1. Comparing Deep-SE cross-project SP estimation replication results (Rep) to the original study results (Orig) [2], and to the baselines. The results of the Wilcoxon test ($\hat{A}_{12}$ effect size in parentheses) for Deep-SE (Rep) vs. Mean and Median baselines are shown in the last column. The best results per project (among all approaches except Deep-SE (Orig)) are highlighted in bold.

(a) Within-Repository Training

Source	Target	Method	MAE	MdAE	SA	Deep-SE (Rep) vs.
		Deep-SE (Orig)	1.07	-	-	
MESOS	USERGRID	Deep-SE (Rep)	1.16	0.96	39.84	
(ME)	(UG)	Mean	1.02	0.19	46.99	1.000 (0.42)
		Median	**0.89**	**0.00**	**54.10**	1.000 (0.37)
		Deep-SE (Orig)	1.14	-	-	
USERGRID	MESOS	Deep-SE (Rep)	1.51	1.01	16.18	
(UG)	(ME)	Mean	1.52	**0.80**	15.57	0.282 (0.51)
		Median	**1.50**	1.00	**16.27**	0.802 (0.49)
		Deep-SE (Orig)	2.75	-	-	
TISTUD	APSTUD	Deep-SE (Rep)	4.37	2.98	12.99	
(AS)	(AP)	Mean	**4.27**	**2.18**	**15.05**	0.918 (0.48)
		Median	4.38	3.00	12.73	0.573 (0.50)
		Deep-SE (Orig)	1.99	-	-	
TISTUD	TIMOB	Deep-SE (Rep)	3.38	2.39	20.31	
(AS)	(TI)	Mean	3.45	2.82	18.69	<0.001 (0.54)
		Median	**3.17**	**2.00**	**25.24**	1.000 (0.45)
		Deep-SE (Orig)	2.85	-	-	
APSTUD	TISTUD	Deep-SE (Rep)	**2.70**	**2.07**	**48.25**	
(AP)	(AS)	Mean	3.38	3.24	35.20	<0.001 (0.59)
		Median	3.17	3.00	39.30	<0.001 (0.56)
		Deep-SE (Orig)	3.41	-	-	
APSTUD	TIMOB	Deep-SE (Rep)	**3.51**	**2.53**	**40.34**	
(AP)	(TI)	Mean	4.36	4.24	25.78	<0.001 (0.64)
		Median	4.19	4.00	28.67	<0.001 (0.62)
		Deep-SE (Orig)	3.14	-	-	
MULE	MULESTUDIO	Deep-SE (Rep)	3.64	**2.04**	17.61	
(MU)	(MS)	Mean	3.34	2.71	24.42	0.775 (0.49)
		Median	**3.26**	3.00	**26.26**	0.997 (0.46)
		Deep-SE (Orig)	2.31	-	-	
MULESTUDIO	MULE	Deep-SE (Rep)	2.77	2.47	36.28	
(MS)	(MU)	Mean	3.05	**1.77**	29.83	0.004 (0.54)
		Median	**2.60**	3.00	**40.24**	0.997 (0.46)

(b) Cross-Repository Training

Source	Target	Method	MAE	MdAE	SA	Deep-SE (Rep) vs.
		Deep-SE (Orig)	1.57	-	-	
TISTUD	USERGRID	Deep-SE (Rep)	3.47	3.50	0.98	
(AS)	(UG)	Mean	3.08	2.82	12.02	1.000 (0.42)
		Median	**2.30**	**2.00**	**34.40**	1.000 (0.27)
		Deep-SE (Orig)	2.08	-	-	
TISTUD	MESOS	Deep-SE (Rep)	3.18	3.20	14.70	
(AS)	(ME)	Mean	3.28	**2.82**	11.95	0.011 (0.52)
		Median	**2.58**	3.00	**30.85**	1.000 (0.39)
		Deep-SE (Orig)	5.37	-	-	
MDL	APSTUD	Deep-SE (Rep)	5.03	3.77	63.29	
(MD)	(AP)	Mean	9.84	8.95	28.21	<0.001 (0.81)
		Median	**3.97**	**3.00**	**71.05**	1.000 (0.43)
		Deep-SE (Orig)	6.36	-	-	
MDL	TIMOB	Deep-SE	**3.34**	**1.96**	**75.82**	
(MD)	(TI)	Mean	11.19	11.95	18.91	<0.001 (0.92)
		Median	4.19	4.00	69.63	<0.001 (0.63)
		Deep-SE (Orig)	5.55	-	-	
MDL	TISTUD	Deep-SE (Rep)	**2.64**	**1.66**	**80.35**	
(MD)	(AS)	Mean	11.45	11.95	14.90	<0.001 (0.97)
		Median	3.17	3.00	76.44	<0.001 (0.58)
		Deep-SE (Orig)	2.67	-	-	
DM	TIMOB	Deep-SE (Rep)	3.81	2.65	59.52	
	(TI)	Mean	5.61	5.03	40.35	<0.001 (0.72)
		Median	**3.46**	**1.00**	**63.22**	1.000 (0.45)
		Deep-SE (Orig)	4.24	-	-	
USERGRID	MULESTUDIO	Deep-SE (Rep)	3.95	2.11	6.18	
(UG)	(MS)	Mean	4.04	2.20	4.00	0.008 (0.54)
		Median	**3.91**	**2.00**	**7.16**	0.917 (0.48)
		Deep-SE (Orig)	2.70	-	-	
MESOS	MULE	Deep-SE (Rep)	3.20	2.31	7.37	
(ME)	(MU)	Mean	**2.89**	**1.81**	**16.56**	0.999 (0.46)
		Median	2.92	2.00	15.65	1.000 (0.45)

Table 5: RQ3.2. Comparing the cross-project prediction accuracy (in terms of MAE, MdAE, SA) of Deep-SE and the baselines Mean and Median. The last column shows the result of the Wilcoxon statistical test ($\hat{A}_{12}$ Effect size in parentheses) for Deep-SE. The best results for each project are highlighted in bold.

Source	Target	Method	MAE	MdAE	SA	Deep-SE vs.
APSTUD, TIDOC, TIMOB, TISTUD	ALOY	Deep-SE	2.15	2.09	47.06	
		Mean	2.83	2.91	30.29	<0.001 (0.67)
		Median	**2.10**	**2.00**	**48.29**	0.770 (0.48)
ALOY, APSTUD, TIDOC, TIMOB, TISTUD	CLI	Deep-SE	2.72	2.43	24.35	
		Mean	2.80	2.50	22.05	0.089 (0.53)
		Median	**2.38**	**2.00**	**33.91**	0.940 (0.46)
ALOY, CLI, APSTUD, TIDOC, TIMOB, TISTUD	DAEMON	Deep-SE	2.95	**2.11**	23.87	
		Mean	3.04	2.51	21.34	0.166 (0.53)
		Median	**2.89**	3.00	**25.25**	0.526 (0.50)
APSTUD, TIMOB, TISTUD	TIDOC	Deep-SE	2.53	**1.49**	39.09	
		Mean	2.91	3.15	29.92	<0.001 (0.64)
		Median	**2.45**	2.00	**40.99**	0.006 (0.53)
TDQ, TESB	TDP	Deep-SE	1.60	**1.05**	50.16	
		Mean	2.46	2.29	23.01	<0.001 (0.75)
		Median	**1.53**	2.00	**52.32**	0.021 (0.54)

Table 6: RQ4. Results achieved by Deep-SE on the Tawosi dataset when the training set is augmented by using older issues from the repository that the project belongs to (AUG), compared to Deep-SE's within-project results from RQ1.2 (WP), and to baseline estimators. The best value per project is highlighted in bold.

Project	Method	MAE		MdAE		SA		Project	Method	MAE		MdAE		SA	
		AUG	WP	AUG	WP	AUG	WP			AUG	WP	AUG	WP	AUG	WP
ALOY	Deep-SE	**2.59**	1.51	2.67	**1.28**	**31.11**	39.67	CONFSERVER	Deep-SE	1.04	**0.91**	**0.65**	**0.64**	63.71	**52.28**
	Mean	3.13	2.23	**2.62**	2.17	16.61	10.84		Mean	1.95	1.35	2.14	1.45	32.16	29.17
	Median	2.80	**1.44**	3.00	2.00	25.47	**42.53**		Median	**0.96**	0.96	1.00	1.00	**66.67**	49.64
TISTUD	Deep-SE	1.73	1.63	**1.80**	**1.38**	50.19	48.08	FAB	Deep-SE	0.87	0.86	**0.72**	**0.71**	52.91	61.06
	Mean	1.91	2.01	2.33	2.16	45.09	35.93		Mean	1.00	1.19	0.78	1.10	45.95	46.21
	Median	**1.51**	**1.51**	2.00	2.00	**56.54**	**51.89**		Median	**0.67**	**0.67**	1.00	1.00	**64.07**	**69.76**
APSTUD	Deep-SE	4.49	4.31	2.83	2.70	25.09	27.37	STL	Deep-SE	1.10	1.18	**0.88**	1.12	36.73	1.91
	Mean	**4.02**	4.00	**2.49**	**2.49**	**33.04**	32.72		Mean	1.24	0.97	1.53	1.02	28.44	19.32
	Median	4.02	**3.99**	3.00	3.00	32.95	**32.81**		Median	**0.95**	**0.95**	1.00	**1.00**	**44.92**	**20.41**
CLI	Deep-SE	**2.04**	**1.76**	**1.94**	**1.30**	**46.62**	**33.44**	COMPASS	Deep-SE	**1.43**	1.63	1.27	**1.34**	**35.53**	15.25
	Mean	3.19	2.14	3.22	2.61	16.49	18.93		Mean	1.89	1.48	1.12	1.63	14.59	23.05
	Median	2.98	1.77	3.00	2.00	21.80	33.04		Median	1.81	**1.38**	**1.00**	2.00	18.31	**28.54**
DAEMON	Deep-SE	3.10	3.29	**2.71**	**2.00**	21.10	20.55	SERVER	Deep-SE	**0.83**	0.89	**0.55**	**0.71**	**36.24**	57.60
	Mean	2.76	2.75	3.05	2.75	29.55	33.53		Mean	0.85	1.56	1.00	1.86	34.23	25.99
	Median	**2.74**	**2.74**	3.00	3.00	**30.21**	**33.81**		Median	0.85	**0.85**	1.00	1.00	34.26	**59.47**
TIDOC	Deep-SE	**3.35**	**2.72**	3.03	1.19	**16.28**	**25.35**	EVG	Deep-SE	0.68	**0.63**	**0.58**	**0.54**	39.21	**19.39**
	Mean	3.47	2.99	3.14	2.59	13.36	18.00		Mean	0.66	0.68	0.91	0.56	41.29	12.99
	Median	3.42	2.77	**3.00**	**1.00**	14.49	24.03		Median	**0.65**	0.69	1.00	1.00	**41.63**	10.68
TIMOB	Deep-SE	**2.46**	**2.41**	**1.87**	**1.81**	**33.79**	**33.90**	TDP	Deep-SE	**1.07**	**0.99**	**0.89**	**0.81**	**62.78**	37.69
	Mean	2.62	2.55	2.18	1.81	29.36	30.23		Mean	2.27	1.17	1.96	1.38	21.27	26.26
	Median	2.53	2.53	2.00	2.00	31.89	30.70		Median	1.47	**0.99**	2.00	1.00	48.97	**37.74**
CLOV	Deep-SE	**3.71**	**3.78**	**1.07**	**1.05**	**42.49**	**47.73**	TDQ	Deep-SE	2.83	**2.47**	2.65	**2.23**	27.82	**49.14**
	Mean	5.35	5.93	4.52	5.30	17.08	18.06		Mean	2.82	4.20	2.81	3.82	28.07	13.65
	Median	4.01	4.01	2.00	2.00	37.83	44.55		Median	**2.22**	2.88	**2.00**	3.00	**43.46**	40.72
CONFCLOUD	Deep-SE	1.48	1.48	1.01	**0.93**	57.04	33.89	TESB	Deep-SE	**1.19**	1.15	**0.77**	**0.73**	**64.16**	21.36
	Mean	2.32	1.49	2.39	1.23	32.82	33.65		Mean	2.52	0.99	2.66	0.99	24.01	32.71
	Median	**1.33**	**1.33**	**1.00**	1.00	**61.54**	**40.87**		Median	1.29	**0.98**	1.00	1.00	61.08	**33.02**

Table 7: Running time (in seconds) for Deep-SE and TF/IDF-SVM on the Tawosi dataset.

Project	Running Time (Seconds)	
	Deep-SE	TF/IDF-SVM
MESOS	844	16
ALOY	144	1
TISTUD	1,641	9
APSTUD	317	3
CLI	226	2
DAEMON	198	1
TIDOC	601	2
TIMOB	2,178	37
CLOV	239	3
CONFCLOUD	179	2
CONFSERVER	303	4
DNN	1,193	7
DURACLOUD	272	1
FAB	214	2
STL	161	2
DM	2,942	43
COMPASS	259	1
SERVER	276	2
EVG	2,421	7
MDL	822	4
MULE	1,586	8
NEXUS	604	7
XD	466	2
TDP	312	2
TDQ	409	2
TESB	466	3
Total	19,273	173

Table 8: Descriptive statistics of title and description text length (in characters) with respect to issue type in each of the projects in the Tawosi dataset used in this study.

Project	Type	# Issues	Title Length (characters)					Description Length (characters)				
			Min	Max	Mean	Median	StD	Min	Max	Mean	Median	StD
MESOS	Bug	722	18	132	61.45	59	18.62	32	200737	6170.51	758.5	18408.78
	Documentation	57	20	87	48.14	44	16.04	40	883	310.91	263	220.84
	Epic	5	25	60	37.20	34	13.52	275	1630	836.80	730	493.15
	Improvement	320	20	104	52.96	52	16.21	15	5514	448.44	313.5	463.45
	Story	9	35	91	59.44	56	20.03	143	1845	506.78	333	534.93
	Task	398	18	128	53.01	52	17.73	33	10810	406.49	274	630.95
	Wish	2	41	72	56.50	56.5	21.92	467	583	525.00	525	82.02
ALOY	Bug	123	27	183	65.59	63	21.40	86	17600	1244.63	709	1981.44
	Improvement	82	17	141	53.74	52.5	18.63	60	2883	476.73	360	431.21
	New Feature	25	23	111	49.20	45	17.94	131	1028	463.60	392	281.67
	Story	9	27	84	49.00	44	19.44	129	3893	754.44	316	1199.66
	Sub-task	2	54	55	54.50	54.5	0.71	137	666	401.50	401.5	374.06
TISTUD	Bug	1521	14	174	72.43	71	23.85	48	133959	1273.97	517	5254.23
	Epic	13	26	86	48.08	45	20.22	74	1986	635.54	425	557.79
	Improvement	537	19	187	65.31	63	21.47	16	28379	648.18	370	1889.44
	New Feature	33	16	127	60.67	62	25.95	58	644	332.88	324	166.99
	Story	495	16	174	59.24	56	21.78	22	4804	370.15	287	363.65
	Sub-task	12	42	105	65.92	60	18.55	42	370	171.83	148	111.19
	Technical task	183	26	156	67.99	65	21.34	10	2415	306.77	249	248.08
APSTUD	Bug	271	20	140	59.70	56	21.90	39	1391584	6582.22	518	84484.66
	Epic	6	8	76	36.83	37.5	24.29	13	1128	462.83	422	427.55
	Improvement	83	20	120	55.66	50	21.62	110	3172	541.08	361	521.62
	Story	80	12	117	53.49	49	19.95	32	3841	471.35	340.5	519.67
	Technical task	36	22	100	50.72	49.5	16.21	42	15902	720.97	187.5	2617.85
CLI	Bug	180	24	142	64.38	60	25.74	51	38579	1655.96	796	3587.80
	Improvement	85	22	112	52.35	50	19.78	31	5794	586.38	369	740.97
	New Feature	11	30	74	51.36	52	13.83	115	982	516.00	508	292.98
	Story	17	18	87	43.06	35	21.91	1	965	359.53	278	295.13
DAEMON	Bug	87	19	132	61.77	59	20.74	45	15949	1102.02	408	2345.84
	Improvement	67	19	87	51.16	49	13.66	1	1209	306.18	276	228.59
	New Feature	40	9	66	44.03	45	13.80	50	2155	407.65	352.5	378.13
	Story	11	17	83	49.36	52	21.35	94	1373	412.45	287	356.25
TIDOC	Bug	339	11	144	54.48	52	20.47	26	20755	654.99	335	1721.61
	Epic	6	37	60	46.83	46	8.98	55	594	293.00	242.5	214.39
	Improvement	309	20	124	51.70	49	16.73	13	3524	490.21	340	502.21
	New Feature	235	19	106	47.94	48	12.57	18	3960	259.98	58	496.98
	Story	86	22	122	50.41	46	20.99	15	6680	457.78	266.5	761.89
	Sub-task	21	23	44	33.76	33	4.83	155	1523	414.29	257	412.81
	Technical task	9	32	82	43.89	41	15.28	47	2778	612.33	435	840.87
TIMOB	Bug	2572	22	229	66.83	64	21.58	47	471384	1747.46	881	9787.44
	Epic	24	17	66	38.21	38	11.04	32	3538	744.58	324.5	941.06
	Improvement	598	18	126	57.90	55.5	18.68	27	15672	727.87	487	964.72
	New Feature	391	17	114	51.99	50	17.76	12	4445	676.61	444	687.59
	Story	314	22	123	54.25	53	16.98	18	9187	547.45	353.5	869.63
	Sub-task	12	32	77	51.08	50	15.25	60	2377	651.33	288.5	811.65
	Technical task	4	24	35	30.50	31.5	4.80	85	1069	463.25	349.5	447.07
CLOV	Bug	127	21	122	62.68	62	19.04	14	15381	1548.98	680	2331.04
	Sub-task	37	22	79	47.89	50	12.80	17	1568	351.76	313	326.19
	Suggestion	172	15	96	52.38	53	16.33	30	5728	569.15	331.5	695.89
CONFCLOUD	Bug	206	22	192	68.39	64.5	24.34	95	48378	1766.48	745	5459.38
	Suggestion	28	16	120	62.04	54.5	28.13	82	3086	574.25	393.5	581.58
CONFSERVER	Bug	418	21	165	72.99	70	25.42	31	57656	2225.66	1021.5	5302.71
	Suggestion	38	16	120	65.79	64.5	25.89	89	5658	967.08	538	1271.85

Table 9: (Continued from the previous page) Descriptive statistics of title and description text length (in characters) with respect to issue type in each of the projects in the Tawosi dataset used in this study.

Project	Type	# Issues	Title Length (characters)					Description Length (characters)				
			Min	Max	Mean	Median	StD	Min	Max	Mean	Median	StD
DNN	Bug	1391	8	178	62.07	59	22.97	16	31557	877.83	535	1483.19
	Improvement	334	5	135	50.74	48.5	19.86	31	2854	514.19	374.5	439.34
	New Feature	7	25	85	60.29	66	21.37	288	1155	575.14	473	302.34
	Story	157	10	81	34.69	33	13.14	28	3727	553.12	487	492.74
	Sub-task	102	15	116	50.85	49	18.42	32	1067	188.37	145	167.30
	Task	73	22	113	52.25	51	18.52	37	4394	386.68	215	552.46
DURACLOUD	Bug	141	17	167	60.13	56	25.34	26	3182	449.89	349	447.53
	Story	156	15	95	44.70	43	18.37	4	1453	347.38	276.5	269.83
	Task	13	18	73	37.54	38	15.13	82	444	240.15	197	116.42
FAB	Epic	4	13	67	47.75	55.5	23.88	137	1215	627.00	578	450.44
	Story	152	16	214	73.73	67	35.53	46	1935	492.39	426	332.58
	Sub-task	103	17	116	48.15	44	17.13	38	665	201.66	157	142.07
	Task	44	16	139	46.36	40	23.58	26	1256	298.57	212	286.92
STL	Bug	26	39	97	62.46	60	16.48	121	20297	1252.00	358.5	3904.90
	Improvement	108	15	99	49.20	51	14.68	10	1402	285.56	218	233.29
	New Feature	48	15	87	50.46	53	15.40	36	1992	338.73	193	384.30
	Task	24	23	88	49.92	49	16.64	34	10322	742.29	231	2076.62
DM	Bug	851	10	115	51.59	50	17.07	33	126922	979.99	420	4514.24
	Improvement	477	17	127	51.88	49	17.32	17	11265	499.31	328	723.75
	Epic	150	8	78	39.24	36	15.43	18	1015	221.33	171.5	198.59
	Story	3865	7	123	46.42	45	16.13	4	15399	407.13	226	749.39
	Technical task	38	13	98	50.68	48	19.98	63	1355	312.79	190	312.35
COMPASS	Bug	8	31	78	48.25	39.5	17.72	83	716	389.75	387.5	217.50
	New Feature	7	17	58	42.29	45	12.98	76	875	331.71	294	263.35
	Story	24	10	95	42.63	41	25.64	95	4002	747.21	615.5	801.17
	Task	221	12	90	42.02	41	13.78	18	2585	265.29	161	357.89
SERVER	Bug	145	18	130	67.41	64	21.04	41	8607	910.28	602	1046.54
	Improvement	181	26	132	58.58	55	19.11	56	5931	613.48	454	691.46
	New Feature	44	25	77	51.66	54	13.44	51	3826	739.82	528.5	817.51
	Task	149	9	157	63.57	64	22.25	18	5286	531.42	339	647.99
EVG	Bug	957	9	185	53.28	51	19.30	12	10086	543.49	297	952.27
	Build Failure	127	10	252	52.88	31	48.83	12	5721	771.61	461	825.09
	Improvement	564	15	134	50.76	49	19.48	14	8509	298.02	178.5	539.83
	Incident	7	23	71	43.29	38	16.34	221	5683	1362.00	493	1972.85
	New Feature	445	7	129	46.75	45	18.96	8	22470	305.10	153	1132.33
	Problem Ticket	31	20	92	53.00	52	19.43	83	1931	597.48	371	475.87
	Question	8	29	90	59.75	56.5	20.33	48	1267	452.88	429.5	380.83
	Sub-task	7	11	48	26.00	25	12.88	55	230	108.71	80	59.88
	Task	678	11	109	46.31	45	16.81	5	5374	213.67	143.5	281.54
MDL	Bug	689	18	141	60.01	59	20.18	12	6742	607.45	418	710.07
	Epic	6	15	61	40.50	39.5	16.79	210	2271	804.50	565	749.27
	Improvement	250	15	158	54.66	51	20.82	8	12764	531.14	264.5	954.89
	New Feature	99	23	174	55.63	51	24.58	27	3302	414.85	231	569.16
	Sub-task	111	17	115	50.63	49	17.65	23	3216	306.04	190	387.91
	Task	239	20	136	52.82	50	20.46	13	4081	544.13	270	701.39
MULE	Bug	1276	19	177	63.85	61.5	21.10	1	36195	981.59	323	2776.85
	Enhancement Request	731	20	134	55.24	53	19.25	1	29539	420.07	222	1428.79
	Epic	3	8	42	23.67	21	17.16	60	368	167.67	75	173.66
	Story	45	19	115	46.18	41	20.77	32	843	172.69	129	155.36
	Task	880	12	147	51.52	49	20.70	1	15031	240.40	128	630.92

Table 10: (Continued from the previous page) Descriptive statistics of title and description text length (in characters) with respect to issue type in each of the projects in the Tawosi dataset used in this study.

Project	Type	# Issues	Title Length (characters)					Description Length (characters)				
			Min	Max	Mean	Median	StD	Min	Max	Mean	Median	StD
NEXUS	Bug	1060	12	163	67.31	64	23.91	45	103960	3223.51	848.5	6878.70
	Improvement	250	19	150	57.87	55.5	20.84	18	18047	755.82	413	1678.51
	Story	86	14	148	53.60	47	26.69	53	2874	652.99	528.5	552.55
	Task	20	22	106	48.45	45	19.76	94	824	296.30	190.5	205.85
	Technical Debt	9	19	69	40.78	46	17.36	30	1052	350.56	325	317.52
XD	Bug	220	20	124	54.11	53	17.72	12	72853	2659.24	678.5	6293.92
	Epic	1	37	37	37.00	37	-	228	228	228.00	228	-
	Improvement	100	17	107	50.89	48.5	20.19	33	9044	498.49	289.5	950.97
	Story	470	13	195	53.06	52	21.04	15	18055	349.87	199	900.43
	Technical task	20	18	94	46.15	39	23.14	38	10747	772.40	202	2359.69
TDP	Backlog Task	5	23	60	43.00	39	14.78	16	988	504.00	646	414.77
	Bug	13	20	121	53.92	43	30.22	127	1985	619.23	489	505.43
	New Feature	86	13	123	49.53	44.5	22.24	35	9649	1138.69	671	1361.68
	Work Item	367	8	136	41.78	40	15.98	12	22222	910.39	489	1789.74
TDQ	Backlog Task	5	49	79	66.60	66	11.24	19	9534	2177.40	256	4125.52
	Bug	577	17	178	70.52	67	26.75	15	29292	966.82	308	2032.50
	New Feature	138	15	112	54.70	52.5	21.59	12	10666	941.93	334	1606.47
	Work Item	139	8	132	56.97	52	26.14	8	8325	400.85	214	907.34
TESB	Bug	283	25	200	71.79	67	27.65	31	63429	1997.04	498	5269.73
	New Feature	420	14	143	52.35	47	22.00	5	5568	501.62	358.5	534.50
	Work Item	27	21	113	51.52	49	23.06	35	6882	528.96	240	1284.71